UNIVERSITY OF CALIFORNIA

Los Angeles

Multiplexed Gene Synthesis in Emulsions

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Chemical Engineering

by

Angus Morgan Sidore

2019

ABSTRACT OF THE DISSERTATION

Multiplexed Gene Synthesis in Emulsions

by

Angus Morgan Sidore
Doctor of Philosophy in Chemical Engineering
University of California, Los Angeles, 2019
Professor Yi Tang, Chair

Improving our ability to build and test DNA sequences will accelerate progress in biology. Multiplexed functional assays (MFAs) can test thousands to millions of DNA sequences for biological function, illuminating comprehensive sequence-function relationships at base-pair resolution. Though transformative, MFAs are currently limited by the sequences that can be built. Natural sequences can be mutagenized, allowing for the generation of all single-amino acid mutants of a particular protein. However, mutagenesis can only explore small subsets of sequence space, far smaller than the typical distance between homologous proteins. Alternatively, small (<200nt) arbitrary DNA sequences can be synthesized as microarray-derived oligo pools for use in MFAs. Unfortunately, sequences over 200nt are difficult to synthesize on microarrays, preventing the generation of protein-length (300-3000nt) libraries. Gene synthesis from microarray-derived oligos is a promising solution to this problem, allowing for the isolated construction and assembly of long DNA sequences. Unfortunately, the current cost of synthesizing genes from microarray-derived oligos is prohibitive, limiting scalability.

In this dissertation, I describe the development of improved methods for multiplexed gene synthesis from microarray-derived oligos. First, I demonstrate the accurate quantification of polymerase error rates and error correction methods in synthetic gene constructs using next-generation sequencing. Next, I describe DropSynth, a low-cost, multiplexed method which builds gene libraries by compartmentalizing and assembling microarray-derived oligos in vortexed emulsions. Finally, I optimize polymerase choice, add error correction, and increase scale to significantly improve the fidelity and scalability of DropSynth. Taken together, these developments represent a new paradigm for the synthetic construction of gene libraries.

The dissertation of Angus Morgan Sidore is approved.

Yvonne Y. Chen

Sriram Kosuri

Roy Wollman

Yi Tang, Committee Chair

University of California, Los Angeles

2019

To my family.

Contents

List of Figures

List of Tables

Acknowledgments

Writing this thesis would not have been possible without help and support from my family, friends and colleagues. First and foremost, I would like to thank my parents for always believing in me and for encouraging me to pursue knowledge and creativity. Mom, thank you for your weekly calls, your endless support, and for being my bodysurfing partner. Dad, thank you for always being an optimist and for introducing me to music. I'm looking forward to rejoining our weekly jam sessions very soon. I would also like to thank my brother Nick for always being a positive and humorous presence in my life.

In addition, I am deeply thankful for my extended family. In particular, I would like to thank Granddad Dave, who instilled in me from an early age the value of scientific thought. Thank you for the Feynman books, the stories from your Navy/nuclear days, and for always encouraging me to work hard. I would also like to thank Grandma Alyce for her endless love and big hugs, and Grandad Vince for instilling the value of hard work and for being my golf/tennis partner. I am grateful to all my uncles, aunts and cousins, in particular Aunt Sisi, who is a fellow PhD survivor and an ever helpful presence, and Aunt Cyndy, who is always optimistic and encouraging.

I've been fortunate to have several mentors who inspired me to pursue graduate school. Freeman, thanks for introducing me to the highs and lows of research, and for giving me the independence to pursue my own project. Adam, thanks for taking a lowly undergrad into your lab and teaching me how to write a paper. Phil, thank you for introducing me to protein engineering, for taking me on as your first graduate student, for introducing me to LA, and for showing me that professors can be cool. I can't wait to see what you do next at UW Madison.

None of this would have been possible without the two mentors who shaped my graduate research: Sri and Calin. Sri, you took me in as a graduate refugee with no advisor and no hope, and shaped

me into the scientist I am today. You have fundamentally improved my approach to research and to life, and you have given me countless opportunities for which I will forever be thankful. You have left an incredible legacy at UCLA, and I can't wait to see what you accomplish at Octant.

Calin, you are perhaps the smartest, most influential, and most easy-going colleague I have ever had. You have taught me so much about science and life, and I will be forever grateful to have contributed to DropSynth with you. I'm going to miss our lunches at Pollo and curry day, but I am so excited to see you become a brilliant professor at University of Oregon.

I'm very grateful to have worked with so many amazing people in the Kosuri lab. Nate, thank you for your awesome ideas, your computational prowess, and for including me on the error correction project. I can't wait to shred fresh pow, hit the sauna, and eat chorizo with you in the Pinnacle of Innovation. Eric, thank you for being the Arthur Morgan to my John Marston. I'm looking forward to seeing you blossom into a true Oakland Man. Cliff, thank you for our long discussions of science, baseball, and housing markets. I'm very excited to see you do great things at UChicago. Jess, I'm very grateful for your support during this stressful time of job searching and thesis writing. I can't wait to celebrate both of our defenses back-to-back. Guillaume, thank you for exchanging ideas and for reviewing my work. I know you'll become a rockstar postdoc wherever you end up. Kim, thank you for making the Kosuri lab fun, and for knitting beautiful hats for all of us. I can't wait to see you embrace data at Fabfitfun. Christina, thanks for your experimental questions that have made me rethink my own research workflows. I'm looking forward to seeing you thrive in Grace's lab. Rocky, thank you for being the pioneer of the Kosuri lab, and for all of your advice about my job search. I look forward to hanging out with you and Oscar very soon. Hwangbeom, thanks for coordinating Kosuri lab basketball, we all miss you. Rishi, thank you for the job connections and for bringing your exceptional ideas to the lab.

I would also like to acknowledge our Lab Managers/Assistants Danny, Suraj, and Jeff. The lab would be in a sorry state of affairs without your leadership. I would also like to thank Joyce, who has been an incredible RA, DropSynth collaborator, and resident lab DJ. I can't wait to see you further your career at Manus Bio. I would also like to acknowledge the hard-working undergrads who have passed through the lab, including Johnny, Megan, Marcia, and Tripp. Special thanks are in order for the brilliant minds at Octant, including Aaron, Henry, Naomi, Leon, and Grace. Thank

you for tolerating my many MiSeq runs.

Thank you to Professors Yvonne Chen, Yi Tang, and Roy Wollman for serving as my committee members and for their guidance over the course of my studies. I am also thankful for the National Science Foundation for providing my funding for the past 3 years.

I am very thankful for the friends who have supported me outside of the lab. In particular, I would like to thank Kane for being a great chap and a true Serb. I have enjoyed our inside jokes over the years and I am looking forward to seeing where your travels take you next. I would also like to thank Rob for being my close department friend and a great resource. Special thanks are in order for all of my friends in LA and the Bay Area, in particular Sunay, Max, RG, Matt, Tristan, TJ, Devin, Tami, Ben, Al, Charles, Paul, Robles, and the Merino Family.

Jen, you are the most important person in my life and I am so thankful I have you by my side. The past few years have been the greatest of my life, and I'm so glad that we met at LACMA 3.5 years ago. Thank you for listening to my practice talks, encouraging me to do my best, and being my life partner. I am so excited for the next stage of our life together. I love you.

Chapter 2 is a version of the published manuscript: N. B. Lubock, D. Zhang, A. M. Sidore, G. M. Church, and S. Kosuri. "A systematic comparison of enzymatic error-correction methods using deep sequencing," *Nucleic Acids Research*, vol. 45, no. 15, pp. 9206-9217, 2017.

Chapter 3 is a version of the published manuscript: C. Plesa†, A. M. Sidore†, N. B. Lubock, D. Zhang, and S. Kosuri "Multiplexed gene synthesis in emulsions for exploring protein functional landscapes," *Science*, vol. 359, no. 6373, pp. 343-347, 2018.

Chapter 4 is a version of the manuscript: A. M. Sidore†, C. Plesa†, J. A. Samson, and S. Kosuri "DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions," *In preparation.*

Vita

EDUCATION

2014 B.S. (Bioengineering), University of California, Berkeley, Berkeley, California

RESEARCH AND WORK EXPERIENCE

2016-2019 Graduate Researcher, Sri Kosuri Lab, UCLA

2015-2016 Graduate Researcher, Phil Romero Lab, UCLA

2014-2015 Research Associate, Adam Abate Lab, UCSF

AWARDS AND HONORS

2016-2019 National Science Foundation Graduate Research Fellowship

PATENTS

A. R. Abate, F. Lan, S. Lim, and A. M. Sidore, "Microdroplet-Based Multiple Displacement Amplification (MDA) Methods and Related Compositions," U.S. Patent Application No. US20180237836A1

PUBLICATIONS

A. M. Sidore†, C. Plesa†, J. A. Samson, and S. Kosuri, "DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions," *In preparation.*

C. Plesa†, A. M. Sidore†, N. B. Lubock, D. Zhang, and S. Kosuri, "Multiplexed gene synthesis in emulsions for exploring protein functional landscapes," *Science*, vol. 359, no. 6373, pp. 343–347, 2018.

N. B. Lubock, D. Zhang, A. M. Sidore, G. M. Church, and S. Kosuri, "A systematic comparison of error correction enzymes by next-generation sequencing," *Nucleic Acids Research*, vol. 45, no. 15, pp. 9206–9217, 2017.

A. M. Sidore, F. Lan, S. Lim, and A. R. Abate, "Enhanced sequencing coverage with digital droplet multiple displacement amplification," *Nucleic Acids Research*, vol. 44, no. 7, pp. e66, 2016.

Chapter 1

Introduction

1.1 Background

Progress in biology is dictated by our ability to read and write DNA. In the past 20 years, our ability to read, or sequence DNA has dramatically improved due to the development of next-generation sequencing (NGS) platforms [1, 2]. These platforms, which allow for the multiplexed incorporation and detection of nucleotides, are capable of reading billions of DNA sequences simultaneously. With this capability, researchers now contribute over 15 petabases of sequence data per year [3], and have used this information to expand knowledge of human disease [4].

Despite recent developments of reconstructing viral and bacterial genomes [5, 6], our capacity to write, or synthesize DNA has lagged behind sequencing. Current methods for DNA synthesis rely on phosphoramidite chemistry, a dated technique employing individual chemical coupling of nucleic acids. Efforts to assemble longer constructs from sequences synthesized by the phosphoramidite method are expensive and difficult to scale [7, 8]. In order to meet aspirational goals such as synthesizing the complete human genome [9], million-fold improvements to existing DNA synthesis techniques are necessary. Furthermore, our ability to test DNA sequences for biological function hinges on the cost and effort of synthesizing such sequences.

1.2 Oligo Synthesis

DNA is commonly synthesized as oligonucleotides (oligos), short, single-stranded DNA segments under 200nt. Almost all oligos are synthesized using phosphoramidite chemistry originally developed by Marvin Caruthers in the 1980s [10]. This chemistry consists of a four-step cycle in which one base is added per cycle (Fig. 1.1, from Kosuri & Church [7]). The process begins when a dimethoxytrityl (DMT)-protected nucleoside phosphoramidite attached to a solid support is removed with mild acid, exposing the 5'-hydroxyl group for chain elongation. A second DMT-protected phosphoramidite is then coupled with the 5'-hydroxyl of the first phosphoramidite. Optionally, 5'-hydroxyl groups left unreacted from phosphoramidite addition are acetylated, preventing further chain elongation and eliminating many single-base deletions. Finally, the phosphite triester linkage between the two nucleoside phosphoramidites is oxidized, producing the phosphate DNA backbone. The cycle is then repeated, allowing the oligo chain to grow in the 3'-5' direction.

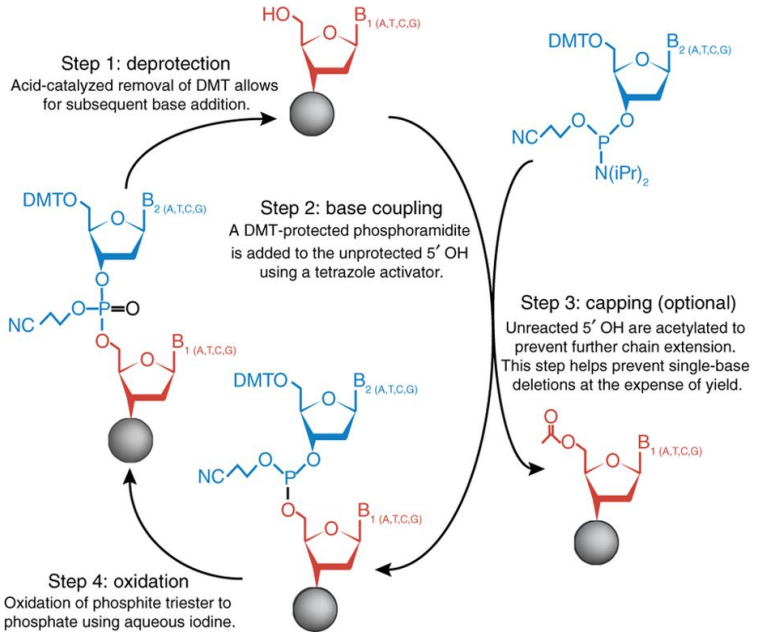

Figure 1.1: **Phosphoramidite method of oligonucleotide synthesis (Kosuri & Church) [7].**

Through a collaboration between Caruthers and Leroy Hood, the first automated oligo synthesizer was built and sold by Applied Biosystems in 1982. Today, the vast majority of DNA synthesis companies use phosphoramidite chemistry. Most of these companies use column-based methods, in which oligos are synthesized in individual columns containing controlled pore glass (CPG) surfaces

(Fig. 1.2A). These synthesizers are capable of producing up to 1536 sequences at nanomolar scales at costs between \$0.05 and \$0.15 per base. Though sequences up to 200nt are possible, larger oligos are difficult to synthesize due to the imperfect efficiency of adding an individual phosphoramidite. For instance, even a 99% coupling efficiency yields a $0.99^{200} = 13\%$ efficiency for a 200mer oligo. Furthermore, longer sequences are more prone to synthesis errors, primarily single-base deletions due to acidic detritylation and inefficiencies in the coupling and capping steps. Despite the advantages of the high efficiency and concentration of synthesized oligos, column-based oligo synthesis methods are often financially impractical for building large libraries (>1000) of sequences.

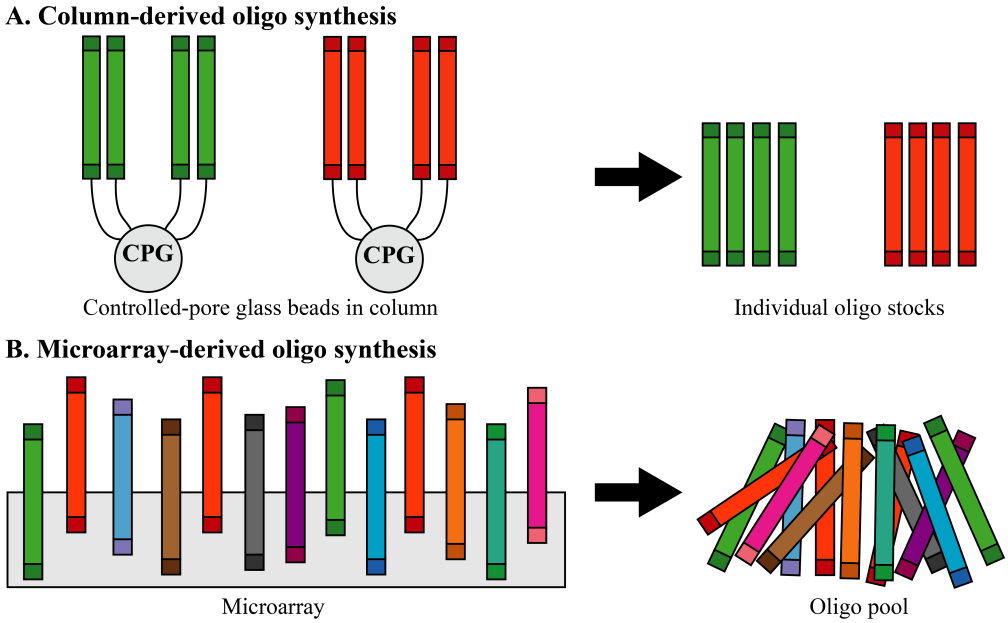

Figure 1.2: **Comparison of column- and microarray-derived oligonucleotide synthesis.** Column-derived oligos are synthesized individually at nanomolar scales, for prices ranging between \$0.05-0.15 per base. Microarray-derived oligos are synthesized in a single pool at femtomolar scales on an arrayed surface, for prices ranging between \$0.00001 to \$0.001 per base.

Oligo synthesis from DNA microarrays is an inexpensive alternative to traditional column-based oligo synthesis. Originally developed for DNA detection, DNA microarrays produce thousands of short DNA strands on chip features using variations of Caruthers' phosphoramidite chemistry. The DNA strands are then cleaved off of the chip, yielding a single oligo pool (Fig. 1.2B). Early techniques, including those developed by Affymetrix in the 1990s, used mask-based procedures to selectively deprotect certain oligos each step using light, allowing for the synthesis of thousands of distinct sequences [11, 12]. Modern maskless techniques, such as the Sureprint technology developed

by Agilent Technologies, use high-definition inkjet printers to deposit precise amounts of each base on a glass slide [13]. These next-generation oligo synthesis methods are capable of producing thousands of <200nt sequences at femtomolar scales, dramatically reducing the quantity of reagents needed per sequence. Consequently, their price ranges from $0.00001 to $0.001 per base, 2-4 orders of magnitude cheaper than column-synthesized oligos. Though array-derived oligos are significantly cheaper, they suffer from a number of disadvantages, including lower fidelity and concentration, spurious depurination, and edge effects due to misalignments of reagent droplets on chip features. Despite these disadvantages, array-derived oligos are an intriguing source of DNA to be used as an input for gene synthesis.

1.3 Gene Synthesis

Because of the inherent limitations of synthesizing oligos over 200nt, alternative methods have been developed to stitch together overlapping groups of oligos into full-length genes. Early developments employed the ligation of partially overlapping adjacent oligos using T4 DNA ligase. These approaches led to the synthesis of the first complete gene, a 77-nucleotide alanine tRNA by Khorana and colleagues [14]. Following the advent of polymerase chain reaction (PCR), a number of ligation-free approaches were developed, including polymerase cycling assembly (PCA) [15]. This method uses a thermostable DNA polymerase to extend overlapping oligonucleotides in a progressive, non exponential manner. More recently, Gibson and colleagues demonstrated the combined use of exonuclease, polymerase and ligase to chew back, anneal and seal overlapping strands of DNA, allowing for the single-step assembly of multiple DNA constructs [16]. In the past decade, optimizations of these approaches have dropped the cost of gene synthesis to under $0.05-0.30 per base. However, the cost of column-based CPG oligo precursors has stagnated, remaining at $0.05-0.15 per base.

Because oligos are the dominant cost of gene synthesis, several groups have developed methods to assemble genes using oligos derived from DNA microarrays. Despite the inherent advantages in cost, DNA microarrays present a number of challenges that must be overcome in order to produce full-length genes. First, because individual oligos exist at femtomolar scales, methods must be developed to amplify them prior to assembly. Second, microarray-derived oligos contain higher error

rates than column-derived oligos, necessitating error correction strategies. Finally, because of the large scale of microarray synthesis (>1000 sequences), cross-hybridization of oligos during assembly becomes a problem [17], limiting both scale and potential applications [7].

A number of recent technologies address these issues. Tian and colleagues were one of the first groups to demonstrate accurate multiplexed gene synthesis, synthesizing 21 genes for the E. coli 30S ribosomal subunit [18]. In order to overcome the hurdles of mishybridization, low concentration and high error, they designed oligos with minimal potential for cross-hybridization, amplified oligos prior to assembly, and employed error correction by hybridization. However, they could only assemble dozens to hundreds of oligos at once, limiting the scalability of their technique. More recently, Quan and colleagues developed a custom inkjet synthesizer that isolates oligos needed for each assembly in individual chambers, amplifies oligos via a single-primer strand displacement amplification, and assembles genes via PCA (Fig. 1.3A) [19]. By physically isolating different groups of oligos, this technique limits the potential for mishybridization of sequences. In a different "off-chip" strategy, Kosuri and colleagues introduced barcoded priming sequences into oligos such that only the oligos needed for a given assembly are amplified together (Fig. 1.3B [20]). These barcoded priming sequences are then digested, and genes are assembled via PCA. By performing a subpool PCR on groups of oligos, this technique simultaneously solves both the oligo concentration and mishybridization problems without the need for specialized chips or synthesizers. These two modern techniques also demonstrated successful error correction of gene assemblies following PCA, further minimizing error rate. Though these two multiplexed gene synthesis approaches are effective, they are very expensive at large scales. The technique developed by Kosuri et. al, for instance, requires the purchase of PCR reagents for every gene assembly needed, which becomes cost-prohibitive for assembling thousands of sequences. In order for large-scale gene synthesis to become widely adopted, future improvements must be made in both cost and effort of assembly.

1.4 Multiplexed Functional Assays

A major goal in synthetic biology is to build and functionally characterize thousands of DNA sequences in a pooled format. These experiments, known as multiplexed functional assays (MFAs)

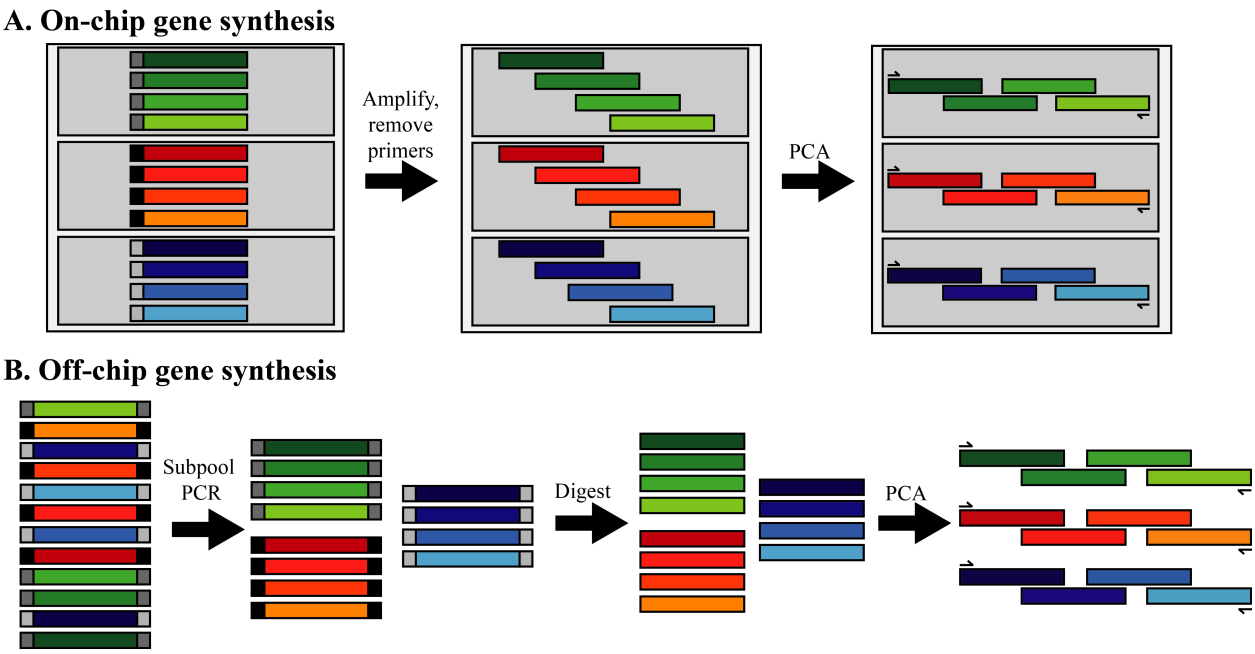

Figure 1.3: **Gene synthesis techniques from microarray-derived oligos**. On-chip methods, first developed by Quan et. al [19], employ specialized DNA microarrays that synthesize, amplify and assemble oligos in separate reaction wells. Off-chip methods, first developed by Kosuri et. al [20],use barcoded primers to separately amplify only those oligos contributing to a given assembly.

can probe proteins and regulatory elements in the form of deep mutational scans [21] and massively parallel reporter assays [22], respectively. An MFA generally consists of five steps, (1) the construction of variant library, (2) the delivery of the library in vivo or in vitro, (3) a functional assay that screens variants by phenotype, (4) next-generation sequencing of variants or barcode identifiers to link sequence to function, and (5) calculation of functional scores for each variant (Fig. 1.4) [23]. The output of such a functional screen is a comprehensive sequence-function map that reveals the fitness effects of many diverse sequences.

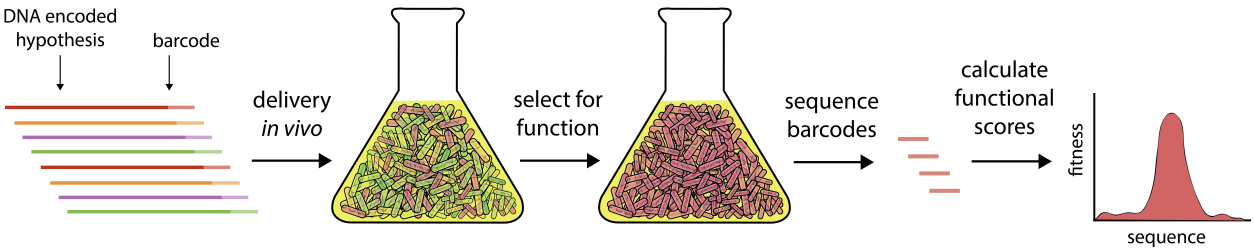

Figure 1.4: **Schematic of a typical multiplexed functional assay (MFA)**. MFAs consist of the construction of a variant library, the delivery of the library, a functional assay that screens variants by phenotype, next-generation sequencing of barcode identifiers to link sequence to function, and the assignment of functional scores to variants [23].

Currently, MFAs are limited by their ability to build or access DNA sequences to test. One existing method, mutagenesis, can create large libraries of sequences with single-base alterations. However, this method is not easily programmable, resulting in uneven distribution of mutations in the resulting sequences. Furthermore, the sequence space explored by mutagenesis is minuscule when compared to the evolutionary distance between two homologous protein sequences. A viable alternative to mutagenesis is the synthetic construction of many sequences. Microarray-derived oligos can be used as libraries [24] but their short lengths (<200nt) limit many applications. Gene synthesis from microarray-derived oligos can produce hundreds to thousands of long-length sequences at relatively low error rates. However, existing multiplexed gene synthesis techniques become cost-prohibitive at large scales. A method of library construction that is simple, cost-effective and scalable will considerable improve our ability to functionally characterize thousands to millions of DNA sequences.

1.5 This Work

In this dissertation we describe methods for improving multiplexed gene synthesis (Chapters 2, 3 & 4). We further show that such methods can be directly inputted into multiplexed functional assays (Chapter 3).

In **Chapter 2**, we develop methods to accurately measure error rates in DNA sequences using NGS. We use these methods to characterize the most commonly used enzymatic error correction methods in gene synthesis, and estimate the error rates of different polymerases.

In **Chapter 3** we introduce a multiplexed gene synthesis method termed DropSynth and use it to synthesize >10,000 genes of up to 669 bp in length. We then test these genes in a multiplexed functional assay and explore the evolutionary and functional landscape of an essential enzyme in *E. coli*.

In **Chapter 4**, we build upon knowledge gained in Chapters 2 & 3 to optimize and improve DropSynth. In particular, we employ polymerase optimization, enzymatic error correction, and increase scale to significantly improve the fidelity and scalability of DropSynth.

References

[1] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nat. Biotechnol.*, vol. 26, pp. 1135–1145, Oct. 2008.

[2] S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nature Methods*, vol. 5, p. 16, dec 2007.

[3] M. C. Schatz and A. M. Phillippy, "The rise of a digital immune system," *GigaScience*, vol. 1, 07 2012.

[4] M. Meyerson, S. Gabriel, and G. Getz, "Advances in understanding cancer genomes through second-generation sequencing," *Nat. Rev. Genet.*, vol. 11, pp. 685–696, Oct. 2010.

[5] H. O. Smith, C. A. Hutchison, 3rd, C. Pfannkoch, and J. C. Venter, "Generating a synthetic genome by whole genome assembly: phix174 bacteriophage from synthetic oligonucleotides," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, pp. 15440–15445, Dec. 2003.

[6] C. A. Hutchison, 3rd, R.-Y. Chuang, V. N. Noskov, N. Assad-Garcia, T. J. Deerinck, M. H. Ellisman, J. Gill, K. Kannan, B. J. Karas, L. Ma, J. F. Pelletier, Z.-Q. Qi, R. A. Richter, E. A. Strychalski, L. Sun, Y. Suzuki, B. Tsvetanova, K. S. Wise, H. O. Smith, J. I. Glass, C. Merryman, D. G. Gibson, and J. C. Venter, "Design and synthesis of a minimal bacterial genome," *Science*, vol. 351, p. aad6253, Mar. 2016.

[7] S. Kosuri and G. M. Church, "Large-scale de novo DNA synthesis: technologies and applications," *Nat. Methods*, vol. 11, pp. 499–507, May 2014.

[8] R. A. Hughes and A. D. Ellington, "Synthetic DNA synthesis and assembly: Putting the synthetic in synthetic biology," *Cold Spring Harb. Perspect. Biol.*, vol. 9, Jan. 2017.

[9] J. D. Boeke, G. Church, A. Hessel, N. J. Kelley, A. Arkin, Y. Cai, R. Carlson, A. Chakravarti, V. W. Cornish, L. Holt, F. J. Isaacs, T. Kuiken, M. Lajoie, T. Lessor, J. Lunshof, M. T. Maurano, L. A. Mitchell, J. Rine, S. Rosser, N. E. Sanjana, P. A. Silver, D. Valle, H. Wang, J. C. Way, and L. Yang, "GENOME ENGINEERING. the genome Project-Write," *Science*, vol. 353, pp. 126–127, July 2016.

[10] S. Beaucage and M. Caruthers, "Deoxynucleoside phosphoramidites—a new class of key intermediates for deoxypolynucleotide synthesis," *Tetrahedron Letters*, vol. 22, no. 20, pp. 1859 – 1862, 1981.

[11] S. P. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas, "Light-directed, spatially addressable parallel chemical synthesis," *Science*, vol. 251, pp. 767–773, Feb. 1991.

[12] A. C. Pease, D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes, and S. P. Fodor, "Light-generated oligonucleotide arrays for rapid DNA sequence analysis," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 91, pp. 5022–5026, May 1994.

[13] E. M. LeProust, B. J. Peck, K. Spirin, H. B. McCuen, B. Moore, E. Namsaraev, and M. H. Caruthers, "Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process," *Nucleic Acids Res.*, vol. 38, pp. 2522–2540, May 2010.

[14] H. G. Khorana, "Total synthesis of the gene for an alanine transfer ribonucleic acid from yeast," *Pure Appl. Chem.*, vol. 25, no. 1, pp. 91–118, 1971.

[15] W. P. Stemmer, A. Crameri, K. D. Ha, T. M. Brennan, and H. L. Heyneker, "Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides," *Gene*, vol. 164, pp. 49–53, Oct. 1995.

[16] D. G. Gibson, "Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides," *Nucleic Acids Res.*, vol. 37, pp. 6984–6990, Nov. 2009.

[17] A. Y. Borovkov, A. V. Loskutov, M. D. Robida, K. M. Day, J. A. Cano, T. Le Olson, H. Patel, K. Brown, P. D. Hunter, and K. F. Sykes, "High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides," *Nucleic Acids Res.*, vol. 38, p. e180, Oct. 2010.

[18] J. Tian, H. Gong, N. Sheng, X. Zhou, E. Gulari, X. Gao, and G. Church, "Accurate multiplex gene synthesis from programmable DNA microchips," *Nature*, vol. 432, pp. 1050–1054, Dec. 2004.

[19] J. Quan, I. Saaem, N. Tang, S. Ma, N. Negre, H. Gong, K. P. White, and J. Tian, "Parallel on-chip gene synthesis and application to optimization of protein expression," *Nat. Biotechnol.*, vol. 29, pp. 449–452, May 2011.

[20] S. Kosuri, N. Eroshenko, E. M. Leproust, M. Super, J. Way, J. B. Li, and G. M. Church, "Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips," *Nat. Biotechnol.*, vol. 28, pp. 1295–1299, Dec. 2010.

[21] D. M. Fowler and S. Fields, "Deep mutational scanning: a new style of protein science," *Nat. Methods*, vol. 11, pp. 801–807, Aug. 2014.

[22] F. Inoue and N. Ahituv, "Decoding enhancers using massively parallel reporter assays," *Genomics*, vol. 106, pp. 159–164, Sept. 2015.

[23] M. Gasperini, L. Starita, and J. Shendure, "The power of multiplexed functional analysis of genetic variants," *Nat. Protoc.*, vol. 11, pp. 1782–1787, Oct. 2016.

[24] G. J. Rocklin, T. M. Chidyausiku, I. Goreshnik, A. Ford, S. Houliston, A. Lemak, L. Carter, R. Ravichandran, V. K. Mulligan, A. Chevalier, C. H. Arrowsmith, and D. Baker, "Global analysis of protein folding using massively parallel design, synthesis, and testing," *Science*, vol. 357, pp. 168–175, July 2017.

Chapter 2

A Systematic Comparison of Error Correction Enzymes by Next-Generation Sequencing

2.1 Abstract

Gene synthesis, the process of assembling gene-length fragments from shorter groups of oligonucleotides (oligos), is becoming an increasingly important tool in molecular and synthetic biology. The length, quality, and cost of gene synthesis are limited by errors produced during oligo synthesis and subsequent assembly. Enzymatic error correction methods are cost-effective means to ameliorate errors in gene synthesis. Previous analyses of these methods relied on cloning and Sanger sequencing to evaluate their efficiencies, limiting quantitative assessment and throughput. Here we develop a method to quantify errors in synthetic DNA by next-generation sequencing. We analyzed errors in a model gene assembly and systematically compared six different error correction enzymes across 11 conditions. We find that ErrASE and T7 Endonuclease I are the most effective at decreasing average error rates (up to 5.8-fold relative to the input), whereas MutS is the best for increasing the number of perfect assemblies (up to 25.2-fold). We are able to quantify differential specificities

This chapter has been published as: N. B. Lubock, D. Zhang, **A. M. Sidore**, G. M. Church, and S. Kosuri. "A systematic comparison of error correction enzymes by next-generation sequencing," *Nucleic Acids Research*, vol. 45, no. 15, pp. 9206-9217, 2017

such as ErrASE preferentially corrects C/G transversions whereas T7 Endonuclease I preferentially corrects A/T transversions. More generally, this experimental and computational pipeline is a fast, scalable, and extensible way to analyze errors in gene assemblies, to profile error correction methods, and to benchmark DNA synthesis methods.

2.2 Introduction

Synthetic DNA is a central tool for biological research [1]. Notably, the initial development of nucleic acid synthesis led directly to the cracking of the genetic code [2]. Today, progress in biology is often limited by the difficulty in producing long, high-quality synthetic DNA [3, 4]. This bottleneck is particularly apparent in the assembly of gene-sized fragments of DNA known as gene synthesis [5].

Currently, gene synthesis relies on the assembly of many oligonucleotides (oligos) of ∼40-150 nucleotide (nt) into a single larger piece of DNA of >1,000 base-pairs (bp) [5]. A variety of methods to assemble oligos into gene-sized fragments exist, but ligation- and polymerase-based assembly methods are the most common [6, 7, 8, 9]. Regardless of the method, the quality of the final product is largely dependent on the quality of the oligos used in the assembly.

Oligos are primarily synthesized using phosphoramidite chemistry first developed by Beaucage and Caruthers in the 1980s [10]. Although these oligos are of high enough quality for common applications such as PCR, their error rates make practical gene synthesis challenging. Several groups have managed to synthesize genes from such oligos, but only find about 5-60% perfect products depending on the size and complexity of the template [11, 12, 13, 14]. This problem is further exacerbated when using lower-cost, but often lower quality oligos from array-based synthesis approaches [15, 16, 17, 18, 19, 20].

Consequently, researchers have developed a number of methods to ameliorate oligo error rate post-synthesis. Size selection methods such as HPLC or PAGE can filter truncated sequences, but are labor-intensive and ineffective against small errors such as single-base deletions, insertions, or substitutions [21, 22]. Hybridization-selection techniques can filter large pools of oligos, but are cost-prohibitive as the number of oligos needed effectively doubles [16, 23]. Sequencing-based retrieval methods can physically pick perfect sequences or separate them by barcoded PCR, but are

time-intensive and can require specialized equipment [24, 25, 26]. Enzymatic error correction is a more commonly-used technique that is relatively inexpensive and effective against most errors. This method employs a variety of different enzymes traditionally used for mutation detection to filter out by binding to or cutting at errors [27, 28, 29, 30].

Two particular classes of proteins are most prevalent in error correction: mismatch binding proteins and mismatch cleaving proteins. Generally, these enzymes recognize distortions in the DNA helix that are caused by mishybridized bases on either strand. In gene synthesis, a pool of perfect and imperfect sequences will be melted and re-annealed pairing perfect and imperfect strands to one another. This produces mishybridized bases that can be recognized by these enzymes. Mismatch binding proteins are used to enrich perfect sequences, while mismatch cleaving proteins are used (often in conjunction with exonuclease trimming) to remove imperfect sequences. The most commonly used mismatch binding protein, MutS, recognizes and binds to all single-base mismatches and a variety of small single stranded loops caused by insertions or deletions (indels) with varying affinity [31, 32, 33, 34, 35]. There are a number of different ways to bind and separate error-containing DNA with MutS including: gel-shift assays, MutS-functionalized columns, and MutS-functionalized magnetic beads [11, 20, 36]. Mismatch cleaving enzymes operate by cutting at or near an error and a variety of different mismatch cleaving enzymes are in use [37]. Broadly, these enzymes can correct errors in two different ways. Similar to mismatch binding methods, perfect sequences can be recovered by filtering them from those cut by mismatch cleaving enzymes. Alternatively, the exonuclease activity is used to trim the error-containing region left over by the mismatch cleaving enzymes. The full length sequences are then recovered by performing a PCR assembly with the trimmed sequences.

Previous assessments of different enzymatic error correction methods have relied on Sanger sequencing of finished gene synthesis products to determine their efficiencies [11, 12, 14, 19, 20]. These studies find that, broadly, the dominant mode of errors in gene synthesis products are single-base deletions and mismatches. However, the prohibitive cost of Sanger sequencing hundreds of thousands of bases has limited the effective characterization and comparison of existing methods. Alternatively, one can turn to the mutation detection literature to find biochemical characterizations of enzymes commonly used in error correction [30, 34, 38, 39, 40]. Although these reports provide

more detailed affinity data, they typically rely on electrophoretic methods and are thus similarly limited in sample size.

In order to overcome these limitations, we developed a custom experimental and computational pipeline that leverages Next-generation Sequencing (NGS) to characterize error rates. Here we report the first in-depth characterization via NGS of both the errors arising from the assembly process, as well as the ability of six of the most commonly used error correction enzymes to eliminate these errors across 11 total conditions. With sample sizes three to four orders of magnitude larger than previous reports, we are able to gain detailed insights into the modality of errors as well as each enzyme's relative ability to correct them. We also used our method to assess the effect of polymerase on assembly quality by comparing a high-fidelity polymerase (Q5) to a low-fidelity one (KAPA2G Robust). We believe that our method can act as a generalizable platform to rapidly and cost-effectively test, characterize, and optimize oligo synthesis parameters or new enzymatic error correction methods.

2.3 Results

Next-generation Sequencing Based Analysis of a Model Gene Assembly

To assess different enzymatic error correction methods, we first constructed a constant reference sequence that served as the base for downstream analyses. We designed this sequence to have a length of 100 bp (not including two 21 bp priming regions for amplification and sequencing), a balanced nucleotide content (26:23:23:28 A:C:G:T content), good coverage of all nucleotide pairs and most triplets (80%) while limiting homo-polymer repeats greater than two, and a 28 bp region in the center that has good melting temperature and low secondary structure to facilitate overlap-extension assembly of the two primers. We assembled this sequence from two 85 nt oligos by a preliminary round of polymerase chain assembly (PCA). We then diluted the products of that reaction and used PCR to amplify the full-length 142 bp construct (Figure 2.1) . We then subject the resulting assembly to multiple rounds of enzymatic error correction and sequence the products at each step.

We expect that errors arising during sequencing will convolute our true signal. In order to limit these errors as much as possible, we developed a stringent data processing pipeline briefly outlined

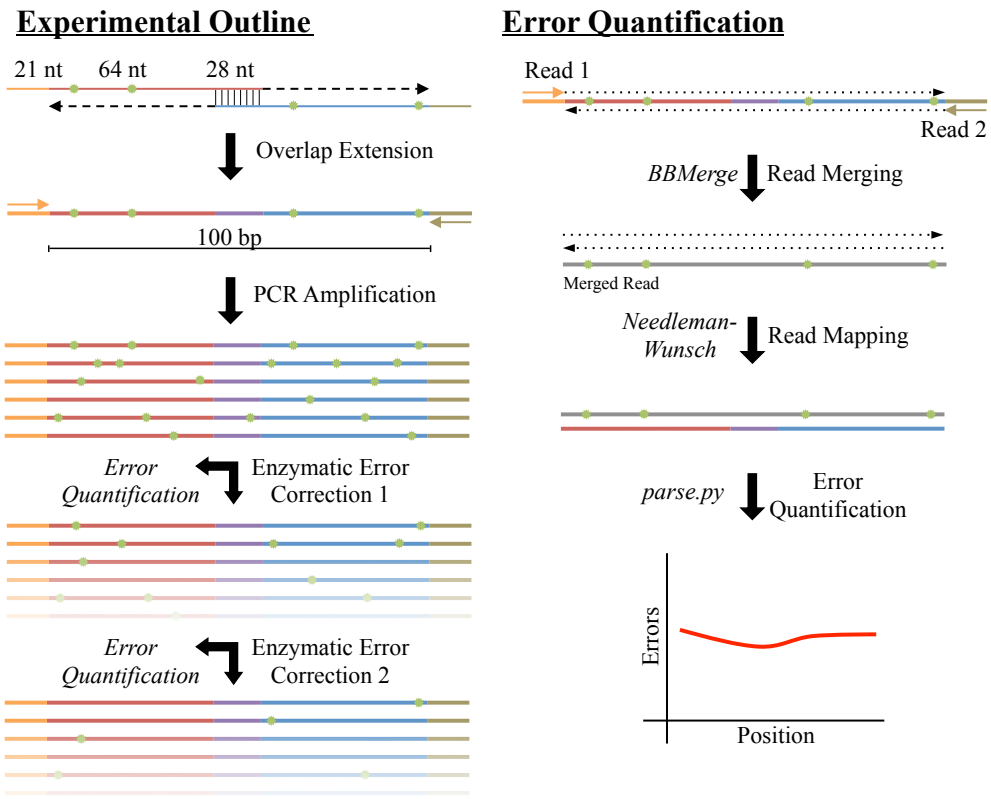

Figure 2.1: **Schematic of Enzymatic Error Correction and Downstream Data Processing.** We assembled our 142 bp product from two 113 nt oligos consisting of a 21 nt primer, a 64 nt payload, and a 28 nt overlap region. After annealing and overlap extension, we amplified our template via PCR, yielding 100 bp of template in-between the primer sites. We then denatured and re-annealed the PCR products to form heteroduplexes, thereby exposing any errors (shown in green). After, we subjected the pool of heteroduplexes to two successive rounds of ten different enzymatic error correction treatments. At each step, we took aliquots and sequenced the products on an Illumina MiSeq with fully overlapping forward and reverse reads. To mitigate sequencing errors, we used BBMerge to merge reads with a perfect agreement between the forward and reverse reads. We then aligned these sequences to the designed reference using an exhaustive Neeleman-Wunsch aligner to minimize alignment artifacts. Finally, we further processed the alignments to quantitate the types and extent of different errors across all conditions.

as follows: First, we cleaned our raw sequencing reads (509,717 per sample on average) by trimming sequencing adapters, removing any reads containg "N" base calls (212 reads on average), and filtering out any reads that aligned to either the PhiX or *E. coli* genomes with `BBDuk` (822 reads on average). This ensures that any spurious reads will not contaminate our alignments and lead to false-positive error calls. Next, we merged our paired end reads together with `BBMerge`, only keeping alignments with perfect correspondence between the forward and reverse reads. Since we sequenced our assembly with fully overlapping reads, each base is effectively sequenced twice. We found that an average of

95.2% of all bases in the merged reads had a Phred33 score (Q) of 41 (∼1/12,600 chance of being miscalled), and 99.8% of all bases on average were above Q30 (1/1000 chance of being miscalled). It should also be noted that most bases were probably above Q41 as this is the default maximum Phred score for most read mergers to maintain backwards compatibility with legacy software. The merging step removed an average of 15.8% of input reads, resulting in an average of 426,514 reads per sample at the end of processing.

After pre-processing the reads, we used a Python implementation of the Needleman-Wunsch aligner, `uta-align`, to align our reads to the perfect reference sequence. We elected to use a Needleman-Wunsch aligner as it is guaranteed to converge on the optimal alignment for a given scoring system [41, 42]. In contrast, typical short read aligners such as `BWA` and `Bowtie2` do not offer such guarantees as they use heuristics to trade accuracy for speed [43, 44]. We find that these heuristics often result in sub-optimal alignments and miscategorization of error sub-types (Figure 2.6, Table 2.2).

Error-doped Oligos Enable Comparisons

In order to assess the sensitivity of our assay, we treated our two-oligo assembly with the error correction cocktail ErrASE and measured the resulting error rates (Figure 2.7). Although we were able to measure significant (Mann-Whitney U, $p << 0.001$, Holm-corrected) reductions in the rate of single-base deletions, multiple-base deletions, and single-base insertions, we were not able to find a significant (Mann-Whitney U, NS, Holm-corrected) reduction between the median rate of mismatches. To ensure that we had a measurable change in error rates for mismatches after enzymatic treatment, we assembled our template from oligos that had errors doped into the sequence. Specifically, we ordered each base with 97% of the intended base, and 1% of the other three nucleotides (not including the 21 bp priming region and the last base of the oligo).

We found that the errors were doped uniformly into our assembly (Figure 2.2A), with the majority of errors being mismatches (90.9%), followed by single base deletions (3.1%), multiple base deletions (2.7%), single base insertions (1.9%), and multiple base insertions (1.5%; Figure 2.2B). Unlike the standard oligo assembly (Figure 2.8), we found no significant difference between the median mismatch rate (3.99×10^{-2}) at any of the four bases (Mann-Whitney U, NS; Figure 2.2C).

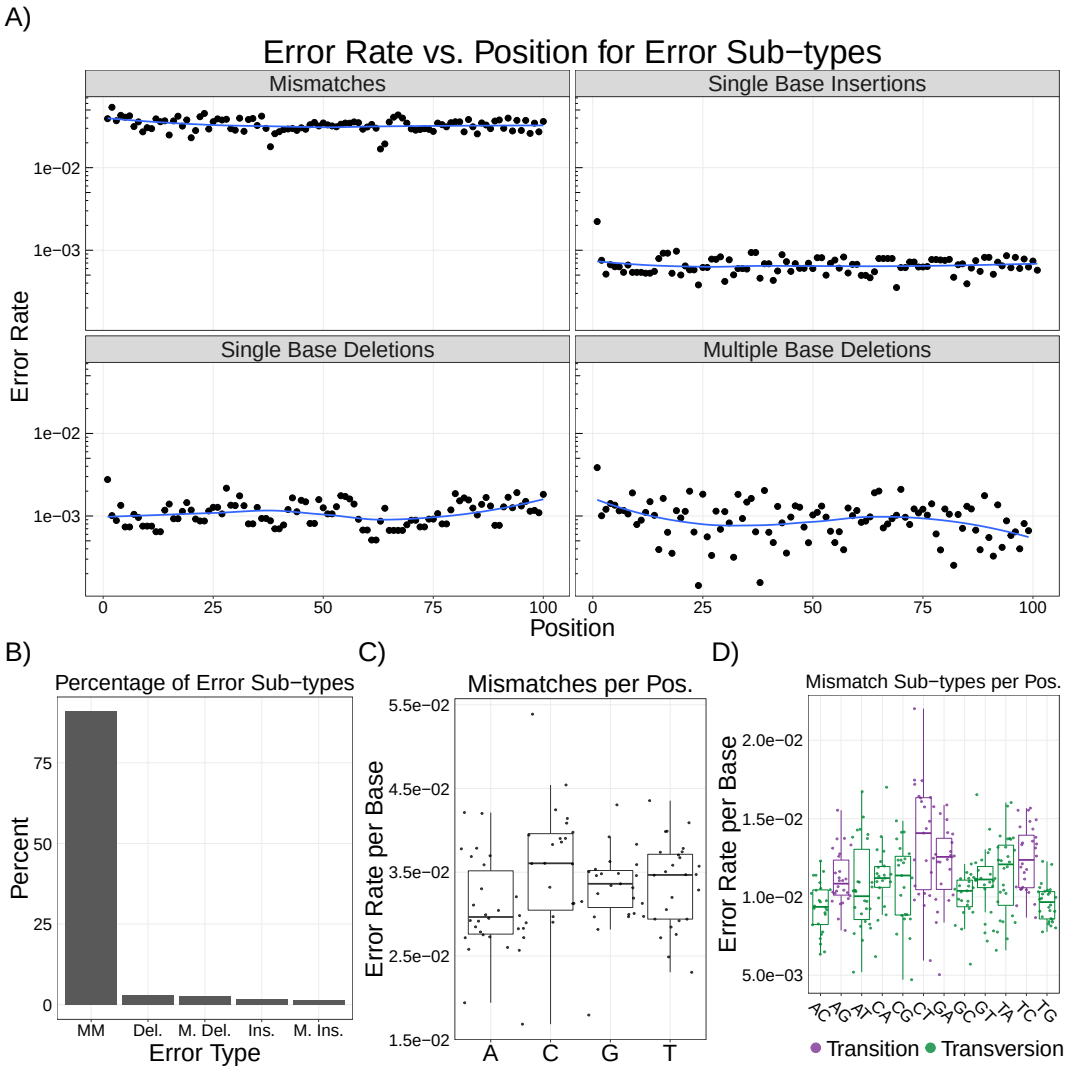

Figure 2.2: **Analysis of Model Gene Assembly Error Rates. A.** The error rates per base are plotted across each position in our model separated by the four major classes of error types. We do not see strong positional effects for errors across the template. **B.** We find a majority of errors on the template are mismatches (MM), followed by single (Del.) and multiple base (M. Del.) deletions; Single (Ins.) and multiple base (M. Ins.) insertions occur at even lower frequencies. **C.** There are no significant differences between the median rate ofmismatches at any base (Mann-Whitney U, NS). **D.** Similarly, there are no significant differences between transitions and transversions (Mann-Whitney U, NS), implying that the errors were doped uniformly into our oligos. **Note:** Blue line is a LOESS fit; box plots are first and third quartile for hinges, median for bar, and 1.5× the inter-quartile range for whiskers.

Similarly, the median rate of individual transitions and transversions were not significantly different from each other (Mann-Whitney U, NS; Figure 2.2D). These data suggest that incorrect bases were doped in to our oligos at an approximately equal rate that exceeded the baseline error rate of KAPA SYBR Fast – the other potential source of mismatches. We note that the median rates of all error

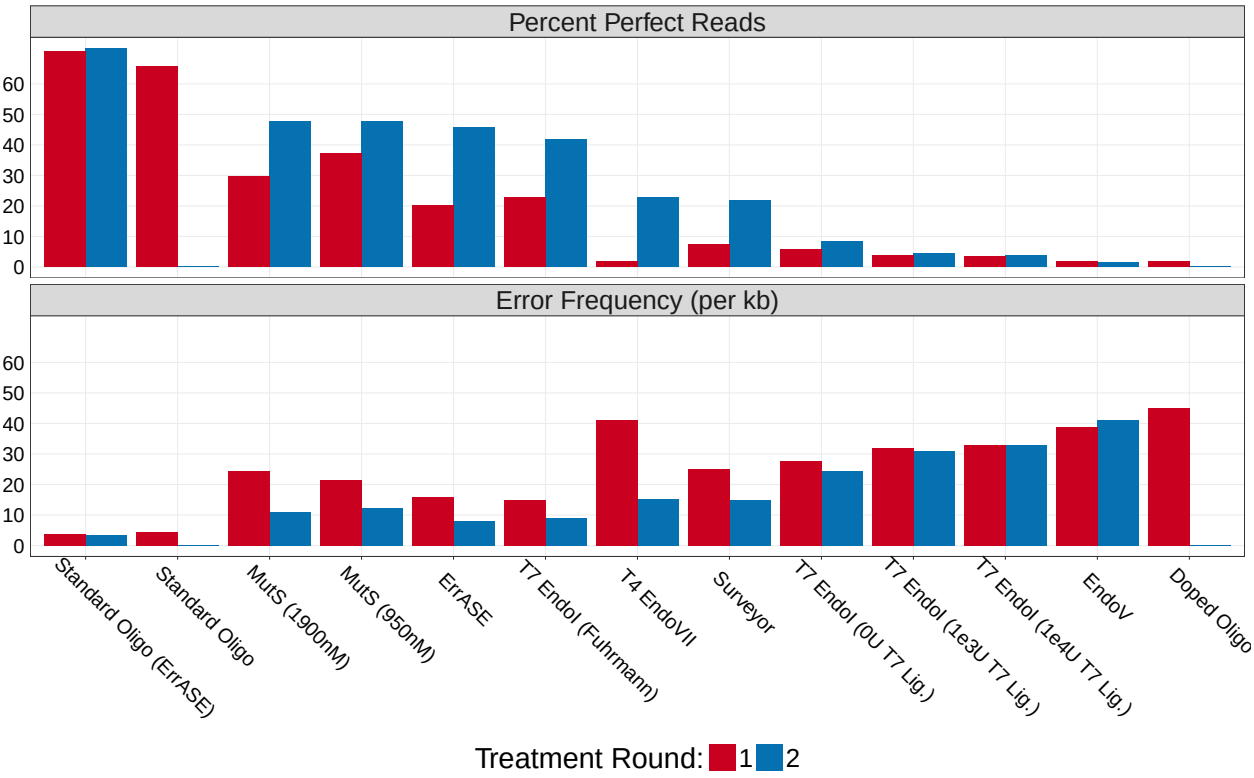

Figure 2.3: **Effectiveness of Enzymatic Error Correction Methods.** Here we compare the error frequency (errors/kb) and number of perfect assemblies for ten different enzymatic error correction methods. We find that MutS is the most effective enzyme at increasing the percentage of perfect assemblies. However, ErrASE is the most effective at decreasing error frequency. Additionally, we see that the efficacy of T7 Endonuclease I is dependent on protocol, and that the addition of a ligase had detrimental effects on sequence quality. **Note:** the x-axis is ordered by decreasing number of perfect assemblies.

types were significantly higher in the error-doped assembly (Table 2.3, Figure 2.9; Mann-Whitney U, $p << 0.001$). Although this is expected for mismatches, we suspect that the higher median error rates for the other error sub-types are a result of the non-standard synthesis required to dope the errors into our oligos.

Enzymatic Error Correction Improves Assembly Quality

Having established the error profile of the error-doped assembly, we evaluated 10 different enzymatic error correction methods using six different enzymes on their ability improve the quality of this assembly (Figure 2.3). As expected, consecutive rounds of enzymatic error correction improved both the relative error frequencies and the number of perfect assemblies. ErrASE was the most effective at decreasing the error frequency, with two rounds of treatment dropping the error frequency

from the doped oligo rate of 45.1 to 7.9 errors/kb. The next most effective enzyme at decreasing error frequency was T7 Endonuclease I (9.1 errors/kb). Based on previous reports in the mutation detection literature, we hypothesized that the addition of a ligase with T7 Endonucluase I would improve correction [39]. We find that the addition of T7 ligase actually decreased assembly quality relative to the no ligase control. In agreement with previous studies, we also find that T7 Endonuclase I is highly sensitive to protocol and concentration as exhibited by the wide range of error frequencies [12, 14]. After T7 Endonuclease I, we found MutS to be the third most effective enzyme at 10.9 errors/kb, with T4 Endonuclease VII, Surveyor, and Endonuclease V following.

However, when looking at number of perfect assemblies sequences, MutS was the most effective enzyme treatment. MutS increased the percentage of perfect sequences in the doped oligo from 1.9% to 47.8% (47.6% for 950nM), while ErrASE increased it to 45.6%, and T7 Endonuclease I increased it to 41.7%. In other words, the oligos that are imperfect after the MutS treatment have more errors on average than those after the T7 Endonuclease I and ErrASE treatments.

Differences in Enzymatic Error Correction

With an average of 426,514 reads per round of error correction, our method provides sample sizes three to four orders of magnitude higher than any previous study. This enabled us to compare the effectiveness of these enzymes on rarer errors such as insertions that would be inadequately sampled with Sanger sequencing. Using the error-doped template as a reference, we measured the relative change in error rates for each position across all different enzymatic error correction methods (Figure 2.4A).

We see that in general, all enzymes tested were able to correct insertions and deletions. We find that enzyme performance (as measured by error frequency or number of perfect assembliess) is directly related to the ability to correct mismatches. For example the best performing enzymes, ErrASE, T7 Endonuclease I, and MutS, were able to decrease the median mismatch error rate relative to the error-doped input by 6.2-, 5.1-, and 4.2-fold, respectively. In contrast, the worst performing enzyme, Endonuclease V, was unable to decrease the median mismatch error rate relative to the error-doped input.

We next sought to measure differences in affinity for specific errors between enzymes (Figures

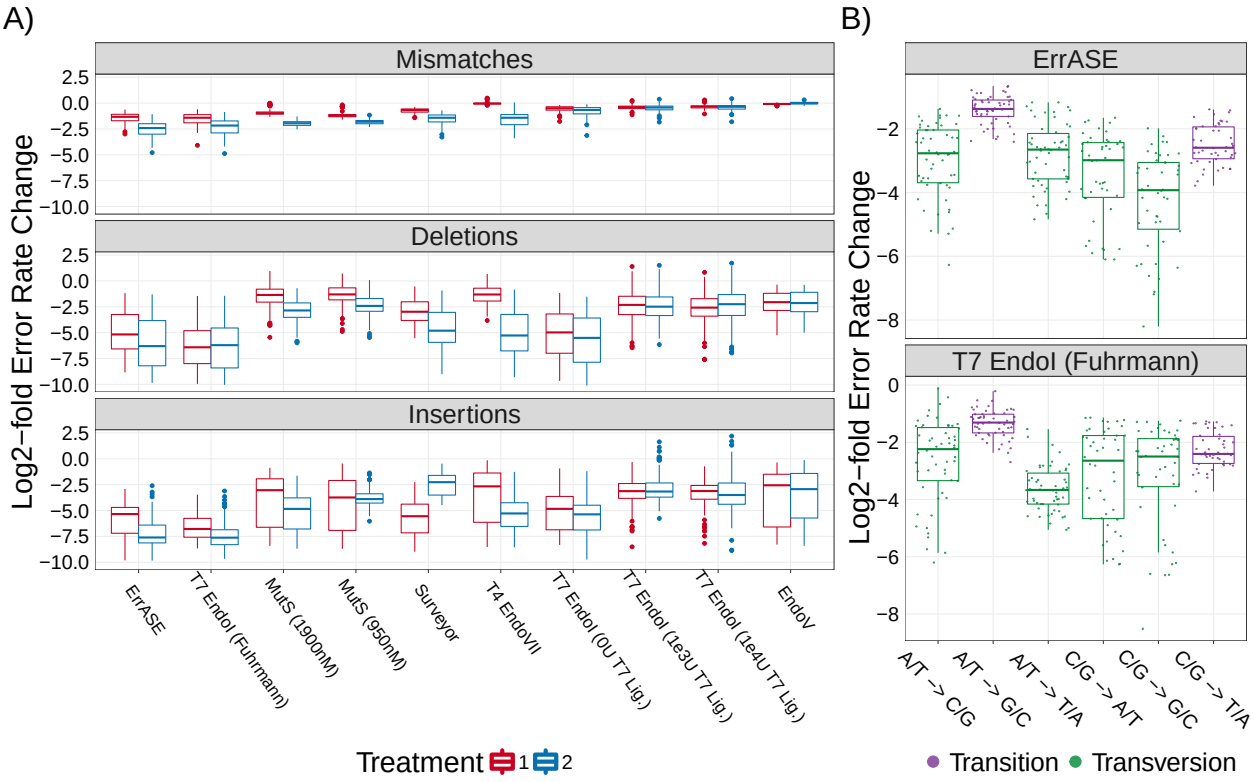

Figure 2.4: **Relative Decrease of Different Error Types. A.** All enzymes were able to correct both single- and multiple-base insertions and deletions. Additionally, we find that the best performing enzymes corrected the highest amount of mismatches. **Note:** the x-axis is ordered by increasing error frequency. **B.** We measure significant differences between the median decrease in C/G → G/C mismatches and the bulk median of all other mismatches after two treatments of ErrASE. Similarly, two treatments of T7 Endonuclease I results in a significant difference between the median decrease in A/T → T/A mismatches compared to the bulk median of all other mismatches (both Mann-Whitney U, $p << 0.001$).

2.10-2.12). We were unable to measure any significant differences between bases for the median fold reduction of insertions and deletions (Kruskal-Wallis, NS) across all enzymes after two treatments. However, we were able to detect significant differences between the median fold reduction of different mismatches (Kruskal-Wallis, $p << 0.001$) across all enzymes after two treatments. Based on these data, we searched for specific mismatch correction biases in our best performing enzymes. For example, we found that two rounds of ErrASE or MutS treatment resulted in a significantly different change in the median fold reduction of C/G → G/C mismatches as compared to the bulk median of all other mismatches (15.2- vs 5.4-fold for ErrASE; 5.1- vs 4.1-fold for MutS; Mann-Whitney U, $p << 0.001$). In contrast, two rounds T7 Endonuclease I did not result in significant changes in the median fold reduction of C/G → G/C mismatches (5.6- vs 5.1-fold; Mann-Whitney U, NS). They did however, significantly change the median fold reduction of A/T → T/A mismatches as compared

to the bulk median of all other mismatches (12.7- vs 4.2-fold; Mann-Whitney U, $p << 0.001$).

Taken together, these data suggest that different enzymatic error correction methods could be used for different applications. For example, GC- or AT-rich constructs would be best corrected by ErrASE and T7 Endonuclease I, respectively. Alternatively, MutS can be used for applications such as protein libraries, where the proportion of perfect sequences are paramount. We also note that the relative rate of correction for transitions and mismatches in general is likely lower than what is measured here due to errors incorporated by the Taq-based KAPA SYBR Fast polymerase during the NGS preparation [45, 46, 47, 48, 49, 50]. For example, the median fold correction of A/T $\rightarrow$ G/C transitions (the most common Taq-based error) was significantly different than that of the bulk median for all other mismatches for ErrASE, MutS, and T7 Endonuclase I (2.6- vs 7.1-fold for ErrASE; 2.8- vs 4.4-fold for MutS; 2.5- vs 6.8-fold for T7 Endonuclease I; Mann-Whitney U, $p << 0.001$).

Analysis of Two Five-oligo Assemblies

In order to investigate the effect of polymerase fidelity on assembly quality, as well as the performance of our method on longer constructs, we assembled two 220-bp constructs from five 60 nt oligos with 20 bp overlaps. To facilitate annealing, we designed the overlap regions to have approximately 50% GC content and minimal secondary structure. We used random nucleotide sequences between the overlap regions with the single restriction being no single nucleotide repeats longer than 4. The resulting nucleotide content of the two constructs are relatively balanced (47:50:62:61 – A:C:G:T for construct one, and 52:53:58:57 – A:C:G:T for construct two). We assembled both constructs with either Q5 or KAPA2G Robust polymerases, and sequenced the assemblies in duplicate with an Illumina MiSeq (~242,000 reads per sample on average after the pipeline filtering). Technical replicates show high correspondence (Figure 2.13) and the error profiles were consistent for each polymerase across the two constructs (Figure 2.14).

As expected, constructs assembled with Q5, a high-fidelity polymerase, had lower error frequencies (2.5 vs 9.7 errors/kb) and a larger percentage of perfect constructs (60.5 vs 10.4%) than KAPA2G Robust, a Taq-based polymerase (Figure 2.5A). The majority of this difference is caused by the higher mismatch frequency in the KAPA2G Robust samples (Table 2.1). The frequencies of errors

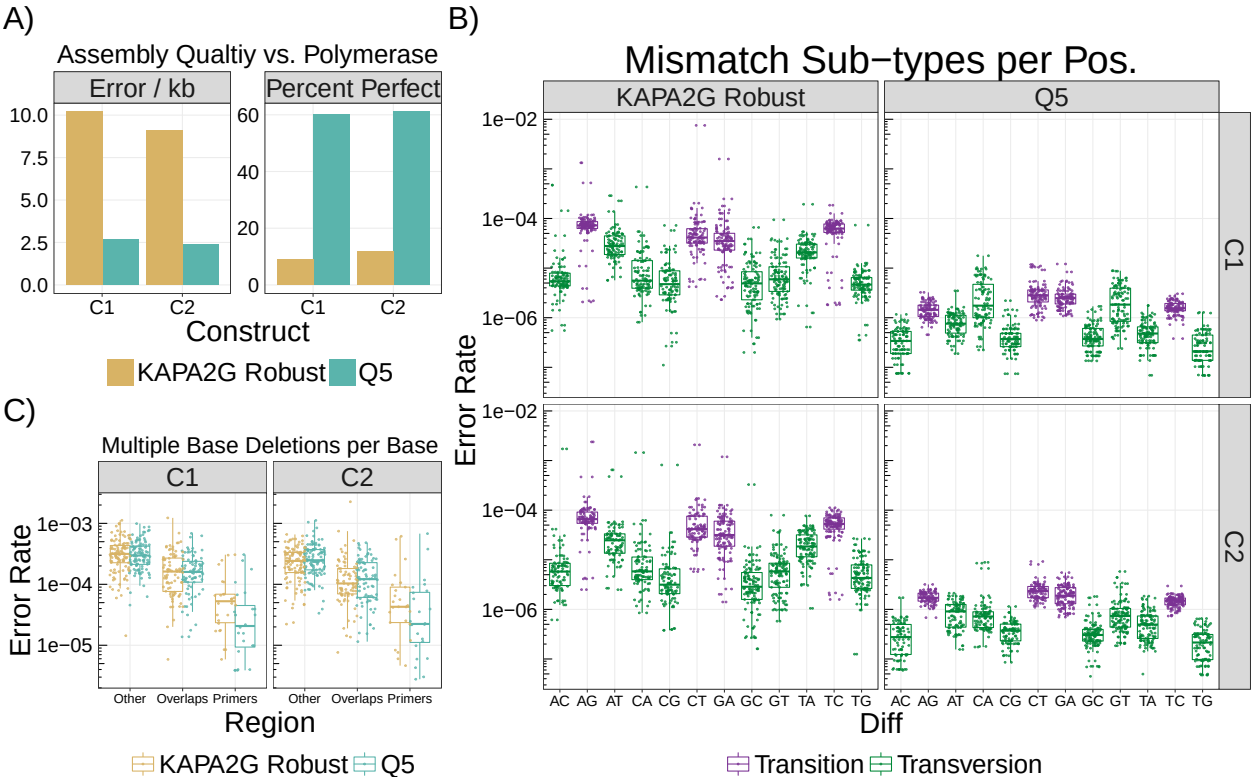

Figure 2.5: **Effect of Polymerase on Assembly Quality.** We assembled two different 220 bp constructs (C1 and C2) from five 60 nt oligos with 20 bp overlaps with Q5 and Taq polymerase. **A.** We used our method to compare the error frequency (errors per kb) and percent perfect assemblies. We see that the average error frequency for both constructs is significantly higher for Taq than for Q5 (9.7 vs 2.5 errors/kb). We observe similar trends for the average percentage of perfect assembiles (60.5% for Q5 and 10.4% for Taq). **B.** Similar to the two-oligo assembly, we find that the Taq-based KAPA2G Robust polymerase also has a higher rate of transitions than transversions (mean of 5.32×10^{-5} vs. 6.40×10^{-6} over both constructs; Mann-Whitney U, $p << 0.001$). **C.** We find that the median rate of multiple base deletions per base in the overlap regions decreased ∼2-fold relative to non-overlapping regions for both polymerases (Mann-Whitney U, $p << 0.001$). Similarly, the median rate of multiple base deltions per base also significantly decreases in the priming regions for both Taq (∼6-fold) and Q5 (∼13-fold) for both constructs (both Mann-Whitney U, $p << 0.001$). The difference in decrease between the polymerases was not significant.

other than mismatches are very similar between the two polymerases (Table 2.1). These errors are likely due to oligonucleotide synthesis, as polymerase and sequencing errors are most often mismatches. Using the previously measured error rates of $\sim 2 \times 10^{-4}$ errors/kb/cycle for Q5, we estimate the expected error frequencies of our assemblies to be ∼0.01 error/kb after 50 rounds of amplification with Q5 polymerase [48]. Since this value is an order of magnitude lower than our measured mismatch rate (0.21 mismatch/kb), we estimate the upper bound of mismatches in oligonucleotide synthesis to be 0.2 mismatches/kb.

In agreement with our two-oligo assemblies (Figure 2.8), the KAPA2G Robust amplified assemblies

Table 2.1: **Estimated error frequencies for five-oligo gene assemblies.** Here, we averaged the errors/kb for both five-oligo assemblies using Q5 and KAPA2G Robust polymerases and their technical replicates across each error type (errors are standard error of the mean). We see that all error subtypes are similar except for mismatches.

Error Type	**Q5**	**KAPA2G Robust**
Mismatches	0.2131 ± 0.0019	7.1388 ± 0.0121
Single Base Deletions	2.0121 ± 0.0062	2.1891 ± 0.008
Single Base Insertions	0.0747 ± 0.0011	0.0816 ± 0.0014
Multiple Base Deletions	0.2326 ± 0.002	0.2342 ± 0.0029
Multiple Base Insertions	0.0014 ± 2e-04	0.0083 ± 4e-04

also had a higher median error rate per base for transitions (5.32×10^{-5}) than for transversions (6.39×10^{-6}) across both constructs (Mann-Whitney U, $p << 0.001$; Figure 2.5B). These errors agree with previous single-molecule studies of this polymerase, and suggest that KAPA SYBR Fast was indeed incorporating mismatches during our NGS preparation for the two-oligo assembly [46, 48]. We note that the KAPA2G Robust assemblies had a very high mismatch rate at the bases immediately before and after the third and fifth overlaps. We did not observe this issue in assemblies of the same oligonucleotide mixtures assembled by Q5.

Next, we measured the effect of the overlapping regions on the number of multiple base deletions (Figure 2.5C). In congruence with our data from the two-oligo assembly, we found that the median rate of multiple base deletions (for a given position in the assembly) was significantly different in the overlap regions than in the rest of the assembly with an average reduction of ∼2-fold for both Q5 and KAPA2G Robust across the constructs (Mann-Whitney U, Holm corrected; $p << 0.001$). We found no significant decrease in the rates of single base deletions in the overlapping regions. Since we added our sequencing primers by annealing to the first and last 15 bp of the constructs, we could also measure the effect of multiple base deletions in the priming region. Again, we found that the rate of multiple base deletions in the priming region was significantly different than both the overlap region and the rest of the assembly, with an average reduction of ∼13-fold for Q5 and ∼6-fold for KAPA2G Robust (Mann-Whitney U, Holm corrected; $p << 0.001$). The differences in reduction between Q5 and KAPA2G Robust were not significant, likely due to a small sample size (n≈25).

2.4 Discussion

One of the most promising methods to improve the quality of gene synthesis products is enzymatic error correction. Previous characterizations of error correction enzymes were limited by Sanger sequencing, which prohibited deep enough sequencing to adequately sample rare variants. Here we surpass this bottleneck by leveraging next-generation sequencing (NGS) and a custom computational pipeline to analyze errors in a model gene assembly. With sample sizes of three to four orders of magnitude greater than any previous study, we were able to accurately quantify error frequencies sample rare errors such as insertions. In addition, NGS precludes the need for time consuming cloning steps. This enabled us to rapidly compare six of the most commonly used error correction enzymes in a total of eleven different conditions in a single experiment, and marks the first comprehensive comparison of enzymatic error correction methods via NGS.

We took multiple steps to minimize the number of false error calls resulting from our method. First, we sequenced our assembly with fully overlapping paired-end reads. Since each base is called independently twice and we only merge reads with a perfect match between the forward and reverse reads, it is unlikely that many sequencing errors made it through this filter. We compared the error profile of the Needleman-Wunsch alignment to two commonly used short-read aligners, `BBMap` and `Bowtie2`. As `BBMap` and `Bowtie2` use heuristics that trade accuracy for speed, we found that their resulting alignments were sub-optimal and led to higher false error calls relative to the Needleman-Wunsch alignment.

We assessed the sensitivity of our method by comparing the error rates of a two-oligo assembly before and after ErrASE treatment. We could measure significant changes in all errors except for mismatches. We hypothesized that our polymerase had re-incorporated mismatches during the NGS preparation. To ensure that we could measure changes in the amount of mismatches, we re-assembled our model sequence with oligos synthesized with 3% of the incorrect base at every position. We expected that the net change in mismatches in the error-doped template after error correction would be larger than the basal error rate of the polymerase, enabling quantification. Additionally, increasing the error rate gives a more realistic number of errors (3-4) per assembly that might occur in a longer gene synthesis.

We then used our method to test the ability of six of the most common error correction enzymes

in eleven total conditions to improve the quality of the error-doped assembly. As expected, we found that all error correction enzymes were able to decrease the error frequency and increase the number of perfect assemblies. We also found that two consecutive treatments of error correction were more effective than one. We then leveraged the large sample sizes generated by NGS to probe specific differences between different enzyme treatments. These data suggest that ErrASE would be the most effective at correcting GC-rich templates, and T7 Endonuclease I is the most effective at correcting AT-rich templates. Alternatively, MutS would be appropriate for the most common applications requiring a single sequenced-verified perfect assembly. The discrepancy of average error frequency and percentage of perfect sequences highlights the importance of using the metrics that are most appropriate for downstream application. In addition, we find that performance of these enzymatic treatments is sensitive to the protocol used as shown in the MutS and T7 Endonuclease I assays.

To test the effect of the polymerase on assembly quality, we assembled two 220 bp constructs from five oligos with both KAPA2G Robust and Q5 polymerases, and compared their error profiles. As expected, we measured a significantly higher number of mismatches in the KAPA2G Robust assemblies than in the Q5. Since the expected mismatch rate of Q5 is lower than our measured value, we estimated an approximate upper bound on the underlying error frequencies of column-synthesized oligos. This is corroborated by the fact that the frequencies of all error types except for mismatches agreed between the two polymerases. Thus, the most common errors in our assemblies were single base deletions, when controlling for polymerase effects. This agrees with previous studies of enzymatic error correction [11, 14, 19]. Two other studies found mismatches to be the most common error. In the first study, this is likely explained by the fact that they amplified their constructs with Taq-polymerase [12]. The second study assembled their genes from chip-synthesized oligos, which might have different error profiles [20]. Lastly, we found that the overlapping regions of our assembly were effective at decreasing the rate of multiple base deletions, but were ineffective for single base deletions.

Our method in its current iteration has limitations. For one, any polymerase misincorporations will convolute the true mismatch correction rate of a given enzyme. While we show that using a high-fidelity polymerase throughout the assembly and NGS library preparation steps ameliorates this issue, we migh still be observing library preparation artifacts. Alternatively, we can incorporate

random barcoding strategies or utilize single molecule sequencing to further eliminate polymerase errors [46, 48, 50]. Second, Illumina sequencing limits our assessments to assemblies < 600 bp. We could extend our methodology to long-read technologies such as PacBio or Oxford Nanopore to assess kilobase-scale gene synthesis products [51]. At these lengths, we would likely have to switch from a Needleman-Wunsch alignment to more optimized versions in order to avoid a significant time penalty [52]. Lastly, our model two-oligo assembly used to analyze enzymatic error correction is not indicative of a typical gene synthesis product as it does not code for a gene, is shorter than standard assemblies (142 bp), is assembled from only two oligos, and has a contrived mismatch error rate.

Overall, our method is a fast and accurate method for looking at errors in arbitrary sequences. We believe that this method will be useful for not only rapidly profiling new enzymatic error correction methods, but for other applications such as assessing the quality of chip-synthesized oligos or developing new gene synthesis methods.

2.5 Materials and Methods

Pre-processing

To ensure that we only analyzed high quality reads, we first ran our sequencing data through a pre-processing pipeline. First, we used `BBDuk` (part of the `BBMap` suite; version 36.14) to trim any Illumina adapters from our reads [53]. Next, we used `BBDuk` to remove any reads with at least 26 bases that match to the PhiX (NC_001422) or *E. coli* (U00096.3) genomes. We also removed any read pairs that had an "N" base call in either one of the reads during this step. We then took the filtered reads and merged read pairs with perfectly overlapping regions with `BBMerge` (also part of the `BBMap` suite; version 36.14) using the `pfilter=1` option.

Alignment and Parsing

After read pre-processing and merging, we use a custom Python script to align our reads to the reference oligo sequence, and parse the resulting alignments to get the positions of all errors. Our Python script uses the `uta-align` (version 0.1.6) package from the Python Package Index (PyPI) to perform a Needleman-Wunsch exhaustive global alignment of the input reads to the reference

sequence [54]. Our script can also provide functionality for performing any alignment supported by the `uta-align` library (e.g. Smith-Waterman local alignments), and allows for tunable gap penalties or match scores.

Once the alignment and parsing is complete, our script will output the results in a tidy csv file with the name of the read, the position of the error, the type of error, and the actual error itself [55]. The types of errors are as follows: M - Mismatch, D - single-base Deletion, I - single-base Insertion, P - multiPle-base deletion, and S - multiple-base inSertion. The errors are classified as: (Original Base)(Mutated Base) for mismatches; the reference base(s) that were deleted for deletions; and the base(s) that were inserted for insertions. Both single and multiple-base insertions are mapped to the "right" of the base in the reference sequence. For example, if the reference sequence was "GATTACA" and we inserted a C at position 3, the resulting alignment can be visualized as:

```
Position:  123-4567
Reference: GAT-TACA
Read:      GATCTACA
CSV:       Read_1, 3, I, C
```

Lastly, if there is a single-base deletion or insertion in a region where there is an identical base adjacent to the mapped position of the error, we distribute the fractional count of the total number of identical bases over each position. For example, if our alignment produced a deletion of A at position 2 in the sequence "TAAAG," our software will note this as a deletion of A at positions 2, 3, and 4, with fractional counts of 1/3 at each of those positions. This compensates for the fact that there are three equally valid alignments in that region.

Error Frequency Calculations and Definitions

To be consistent with previous studies, we calculated the relative error frequency per kb (f) as

$$f = \frac{\sum_{i=1}^{n} x_i \frac{1000}{l_i}}{n} \tag{2.1}$$

where x_i is the number of errors in read i, l_i is the length of that read, and n is the total number of

reads [12]. This is distinct from error rates, which are defined as the number of errors detected at a given base, divided by the total number of sequencing reads in the sample. Error rates can be further separated by the specific error sub-type.

Reagents

All the oligos were synthesized by Integrated DNA Technologies (IDT). The ErrASE Error Correction Kit was purchased from Novici Biotech and is now available as CorrectASE from ThermoFisher. The Surveyor Mutation Detection Kit was from Transgenomic. T4 Endonuclease VII was from Affymetrix. *Thermus aquaticus* MutS DNA mismatch repair protein was from Excellgen. Endonuclease V, T7 Endonuclease I, and T7 DNA Ligase were all from New England Biolabs.

Error-enriched oligonucleotide synthesis and template assembly

The 85-nucleotide (nt) forward and reverse oligos contains 21nt primer sites and 64nt template regions, 63 of which, except for the last base, were doped with 3% errors at each position (Supplementary File 1). This doping is achieved by hand-mixing 1% of every other base into the 97% of the reference base. For example, according to the reference sequence, if a position is supposed to be an A, then 1% of C, T, and G was mixed into 97% A during the initial oligo synthesis by IDT. With 28nt complementary regions, the two oligos were able to anneal and then assembled into a 142-base pair (bp) doubled-stranded template. This template consists of two 21bp primer regions and a 100bp region for error correction and for subsequent next-generation sequencing.

Specifically, to pre-assemble the forward and reverse oligos, 10.4μL nuclease-free water (Ambion), 4μL 5X HF Buffer (New England Biolabs), 0.4μL 25mM dNTP (New England Biolabs), and 0.2μL Phusion High Fidelity Polymerase (New England Biolabs) were added into 5μL 1μM mixed aforementioned forward and reverse oligos. Initially heated at 98C for 30 seconds, the reaction was then cycled 15 times: at 98C for 5 seconds, at 70C for 1 second, ramping down with a speed of 0.5C/second to 50C, at 50C for 30 seconds, and at 72C for 20 seconds. The final extension step was at 72C for 5 minutes. The product after the pre-assembly step was diluted 1:10 in nuclease-free water, 2μL of which, served as template, was added into 35.25μL nuclease-free water, 10μL 5X HF Buffer, 1μL 25mM dNTP, 0.5μL Phusion High Fidelity Polymerase, 1.25μL 10mM mixture of forward (5$'$

TACACGACGCTCTTCCGATCT 3′) and reverse (5′ AGACGTGTGCTCTTCCGATCT 3′) PCR amplification primers to make the total volume of this PCR 50μL (Supplementary File 1). Initially heated at 98C for 30 seconds, the reaction was then cycled 25 times: at 98C for 5 seconds, at 62C for 10 seconds, at 72C for 10 seconds. The final elongation step was at 72C for 5 minutes. Pooled PCR products were then cleaned using QIAquick PCR Purification Kit (Qiagen), and the purified products served as the template for subsequent error correction treatments and sequencing.

Error correction of the synthetic DNA template

ErrASE

Per the manufacturer's instructions, 60μL of $\sim$50ng/μL template in 1X HF Buffer was re-annealed to form heteroduplex by heating at 98C for 1 minute, cooling at 0C for 5 minutes, and incubating at 37C for 5 minutes. Next, 10μL of this re-annealed heteroduplex was added into each well of the 6-well ErrASE tube and was incubated at room temperature for 1 hour. We then combined 2μL from each well as template into the recovery PCR, whose setup and thermocycling conditions were the same as the assembly PCR in the section above. The PCR product using the treated heteroduplex from the first well of the ErrASE tube (presumably has the highest concentration of ErrASE) presented a band, indicating successful recovery after error correction. This product was thus cleaned-up using QIAquick PCR Purification Kit and served as the template for the second iteration of ErrASE treatment.

Surveyor

Per the manufacturer's instructions, $\sim$50ng/μL template in 1X HF Buffer was re-annealed to form heteroduplex by the following thermocycling conditions. First, the sample was heated at 95C for 10 minutes. Then, the temperature was ramped down at 2C/second, and was held at 85C for 1 minute. Finally, the temperature was further cooled down to 25C at 0.3C/second, and was held for 1 minute at every 10C interval. Per Saaem *et al.*, 2μL Surveyor Nuclease S and 1μL Enhancer S were added into 8μL re-annealed heteroduplex [19]. The reaction mixture was then incubated at 42C for 60 minutes. After the treatment was concluded, 2μL of the mixture served as the template in the recovery PCR, whose setup and thermocycling conditions were the same as the assembly PCR.

The product of this recovery PCR, once cleaned-up, entered the next round of Surveyor Nuclease treatment.

Endonuclease V

Similar to Fuhrmann *et al.*, 10μL of $\sim$50ng/μL template in 1X HF Buffer was re-annealed using the cycling condition described in the ErrASE section [12]. We then added 5U of Endonuclease V, 2μL of NEBuffer 4, and nuclease-free water to the re-annealed heteroduplex to make the total volume 20μL. The reaction was incubated at 37C for 24h, and 2μL of this mixture served as the template for the recovery PCR. The cleaned-up product then entered the next iteration of Endonuclease V treatment.

T7 Endonuclease I (Fuhrmann)

As in Fuhrmann *et al.*, 10μL of $\sim$50ng/μL template in 1X HF Buffer was re-annealed using the cycling condition described in the ErrASE section [12]. We combined 2μL of NEBuffer 2, 25U of T7 Endonuclease I, and nuclease-free water to make the final volume 20μL. The reaction was incubated at 37C for 24 hours, and 2μL of the mixture served as the template for the recovery PCR. The cleaned-up product entered the next iteration of T7 Endonuclease I treatment.

T7 Endonuclease I with T7 DNA Ligase

We first re-annealed 100ng of template in 1X HF Buffer according to the ErrASE protocol. Then we combined 2.5μL of T4 DNA Ligase reaction buffer, 10U of T7 Endonuclease I, T7 DNA Ligase (at 0, 1000U, or 10000U), and the appropriate amount of nuclease-free water to make the final volume 25μL. The reaction was then incubated at 25C for 4 hours, and 2μL of the treated sample served as the template for recovery PCR. We used 100ng of the cleaned-up product for the next iteration of T7 Endonuclease I/T7 DNA Ligase treatment.

T4 Endonuclease VII

First, 10μL of $\sim$50ng/μL template in 1X HF Buffer was re-annealed using the cycling condition described in the ErrASE section. Then, 1μL 1M Tris-HCl (pH 8.0), 4μL 50mM $MgCl_2$, 2μL 100mM

β-mercaptoethanol, 1μL 10mg/ml BSA, and 2μL T4 Endo VII (1000U) was added to the 10μL heteroduplex. The reaction mixture was incubated at 37C for 24 hours, and 2μL of which served as the template for the recovery PCR. Then the cleaned-up PCR product entered the next cycle of T4 Endonuclease VII.

MutS

Per the manufacturer's instructions, 250ng/μL in 10mM Tris-HCl (pH=7.8) and 50mM $MgCl_2$ was heated to 95C for 5 minutes followed by cooling at 0.1C/second to 25C. To the re-annealed template, 207.39μL 1X binding buffer (20mM Tris-HCl (pH=7.8), 10mM NaCl, 5mM $MgCl_2$, 1mM Dithiothreitol and 5% glycerol) was added, making the concentration of DNA template to $\sim$11.5ng/μL. This mixture was then aliquoted into two tubes with 109μL in each. Appropriate amount of MutS was added into each of the tubes so that the final MutS concentration was 950nM and 1900nM, respectively. The mixtures were then incubated at room temperature for 20 minutes. Equal volumes of Amylose Resin (New England Biolabs), washed and pre-equilibrated with 1X binding buffer, were added into the tubes. The mixtures were incubated at room temperature for 30 minutes, before being spun down. We purified the supernatants with a Qiagen MinElute kit, and eluted the product in 10μL EB. We used 2μL of the 1:100 diluted elution as the templates for the recovery PCR. Lastly, we pooled the PCR products, cleaned them up, and used them for the next iteration of MutS treatments.

Next-Generation Sequencing using Illumina MiSeq

Each of the control and enzymatically treated samples was prepared as an individual sequencing library. In summary, the sequencing libraries were prepared using two rounds of qPCR, with the first round appending the Illumina P5 sequence and the second appending the P7 sequence as well as the indices. We also note that the KAPA SYBR FAST kit is a Taq-based polymerase. Specifically, the first round of PCR was set up by mixing 25μL KAPA SYBR FAST Universal 2X qPCR Master Mix (KAPA Biosystems), 1μL 10μM Multiplexing PCR Primer 1.0, 1μL 10μM Multiplexing PCR Primer 2.0, 1μL $\sim$100pg/μL error correction DNA template, and 22μL nuclease-free water. Per the manufacturer's instructions, the 2-step thermocycling protocol was used for the qPCR reactions.

Once the signals reached the plateaus, the reactions were stopped and cleaned-up using Agencourt AMPure beads, according to the manufacturer's instructions. The final elution volume was 30μL. To set up the second round of PCR, 25μL KAPA SYBR FAST Universal 2X qPCR Master Mix, 1μL 10μM Multiplexing PCR Primer 1.0, 1μL 10μM PCR Primer each with a distinct index, 1μL $\sim$100pg/μL template from the first round PCR, and 22μL nuclease-free water. The thermocycling and cleaned-up procedures remained the same as those in the first round of PCR. Then, the individually prepared sequencing libraries were quantified using the Library Quantification Kit-Illumina (KAPA Biosystems), according to the provided protocol. Barcoded libraries were subsequently mixed to $\sim$10nM concentration, and the mixed libraries were quantified again before being loaded onto an Illumina MiSeq with a V2 300 cycle kit.

Five-oligo Assembly with High- and Low-fidelity Polymerases

We designed two 220-bp constructs that can be assembled from five 60-nucleotide (nt) oligos each (Supplementary File 1). Each overlap region between adjacent oligos is 20-bp in length, and the first and last oligo contain 15-bp forward and reverse priming regions used for assembly. All overlap and priming sequences were taken from the set designed in Eroshenko *et. al* to minimize cross-hybridization and maximize T_m similarity [56]. Each set of five oligos was synthesized by Integrated DNA Technologies (IDT) with no modifications and pooled into two 1μM five-oligo mixes.

To pre-assemble the five-oligo construct, 5μL of each 1μM five-oligo mix was added to 10μL of NEBNext Q5 HotStart HiFi PCR Master Mix or KAPA2G Robust HotStart ReadyMix and 5μL nuclease-free water. Initially heated at 98C for 30 seconds, the reaction was then cycled 15 times: at 98C for 5 seconds, at 70C for 1 second, ramping down with a speed of 0.5C/second to 50C, at 50C for 30 seconds, and at 72C for 20 seconds. The final extension step was at 72C for 5 minutes. The product after the pre-assembly step was diluted 1:10 in nuclease-free water, 2μL of which, served as template, was added into 20.5μL nuclease-free water, 25μL of Q5 or KAPA2G Robust master mixes, and 1.25μL 10mM mixture of forward and reverse amplification primers flanking the outer oligos of each construct. Initially heated at 98C for 30 seconds, the reaction was then cycled 20 times: at 98C for 5 seconds, at 62C for 10 seconds, at 72C for 10 seconds. The final elongation step was at 72C for 5 minutes. Pooled PCR products were then purified using a DNA Clean and Concentrator-5 (Zymo).

We prepared each assembly as an individual sequencing library with two techincal replicates. The sequencing libraries were prepared using a single round of PCR, which appended both the Illumina P5 and P7 sequences as well as the indices. Specifically, 0.01ng of template was added to 20.5μL nuclease-free water, 25μL Q5 or KAPA2G Robust (depending upon initial condition), and 1.25μL 1μM forward and reverse sequencing primer with corresponding distinct indices. Each library was amplified for a small number of cycles ($\sim$12-14) empirically determined using KAPA SYBR FAST Universal 2X qPCR Master Mix (KAPA Biosystems). We estimate the total number of amplification cycles to be < 50 (< 15 for pre-amplification, 20 for amplification, and 12-14 for NGS prep). Individually prepared sequencing libraries were quantified using an Agilent TapeStation 2200. Barcoded libraries were subsequently pooled and mixed to 20nM concentration, and prepared for sequencing on a 500-cycle V2 MiSeq (Illumina).

2.6 Supplementary Information

Analysis of a Two Oligo Assembly

We applied our pipeline to quantify the different types of errors found in our two-oligo assembly of standard (not error doped) oligos (Figure 2.8). We find that on average about one-third of assemblies contain errors, with an overall error frequency of approximately 4.3 errors per kb. We find that mismatches account for the majority of errors ($\sim$75%), followed by single ($\sim$14%) and multiple-base deletions ($\sim$8%) (Figure 2.8A). The mismatches segregate into two significantly different populations, with the median error rate per base being higher at A's (4.33×10^{-3}) and T's (4.25×10^{-3}) than at G's (1.68×10^{-3}) and C's (1.91×10^{-3}) (Figures SB, C; Mann-Whitney U, $p << 0.001$, Holm-corrected). Furthermore, we find that the median rate of transitions was significantly higher than that of transversions for each base (Figure 2.2C; Mann-Whitney U, $p << 0.001$, Holm-corrected). All of these observations indicate that much of the mismatch error rate is due to polymerase misincorporation during the amplification steps for assembly and sample-preparation for sequencing. Specifically, we used the Taq-based KAPA SYBR Fast polymerase during next-generation sequencing library preparation steps. Consistent with our observations, misincorporations caused by Taq occur most often at A's and T's, and are preferentially A/T $\rightarrow$ G/C transitions (49–52). However, we

cannot completely rule out the effect of errors in the oligo synthesis as our error frequency of 4.3 errors per kb is higher than the ∼3 error/kb expected from 50 rounds of amplification at previously reported Taq error rates (52–54).

Next, we quantified the rates of single- and multiple-base deletions. We find that the median single-base deletion rate per position (5.64×10^{-4}), and that this rate did not vary significantly over the positions (Mann-Whitney U, NS). We also find that multiple-base deletions occur at a similar rate as single-base deletions (3.35×10^{-4}), and measure positional effects for where they occur. Some of this dependence can be explained by the fact that the positions of multiple-base deletions are mapped to the left-most deleted base. Thus, we expect the total number of multiple-base deletions to be highest at position one and decreasing after, since there are the most possible combinations of multiple-base deletions at that position. In addition, we measure a significant decrease in the median multiple-base deletion rate in the annealing region (positions 36-64) of our assembly (Mann-Whitney U, $p << 0.001$). Large deletions in this region would disrupt the hybridization of the initial assembly, leading to sequence drop-outs and a decrease in the measured number of deletions. We also expect the multiple deletion rate to drop towards the end of the sequence due to a "TATATAT" motif at positions 92-98. Any "TA" deletion (or other substring contained multiple times in the motif) will map to the left-most position, 92.

Finally, we quantified single-base insertions. These errors occur at median rate per position of 9.65×10^{-5}) and exhibit no positional dependence besides an outlier at position 1. An incomplete primer trimming by `BBDuk` can explain this outlier. Here, 57 of the 152 single-base insertions are a "T," corresponding to the last base of the primer sequence directly upstream of our first base. Without these 57 bases, the rate of single-base insertions falls closer to the expected median value. Our method is also able to detect multiple-base insertions, which occur at a median rate of 6.16×10^{-6}).

Availability

The computational pipeline described above is open source, free to use under the MIT license, and available at `https://github.com/kosurilab/errorCorrect`. For the final analysis and figure production, we used `R` (version 3.3.*) and `ggplot2` [57, 58].

Accession Numbers

Sequencing data are available from the sequencing read archive (SRA) with the accession number SRP110084.

Funding

This work was funded by the funds from the US Department of Energy [DE-FC02-02ER63421 to S.K.], National Institutes of Health New Innovator Award [DP2GM114829 to S.K.], Searle Scholars Program [to S.K.], Office of Naval Research [N000141010144 to S.K. and G.M.C.] and a Ruth L. Kirschstein National Research Service Award [GM007185 to N.L.].

Conflict of interest statement.

None declared.

Acknowledgments

The authors would like to acknowledge members of the Kosuri Lab for comments on the manuscript, especially Rocky Cheung and Calin Plesa.

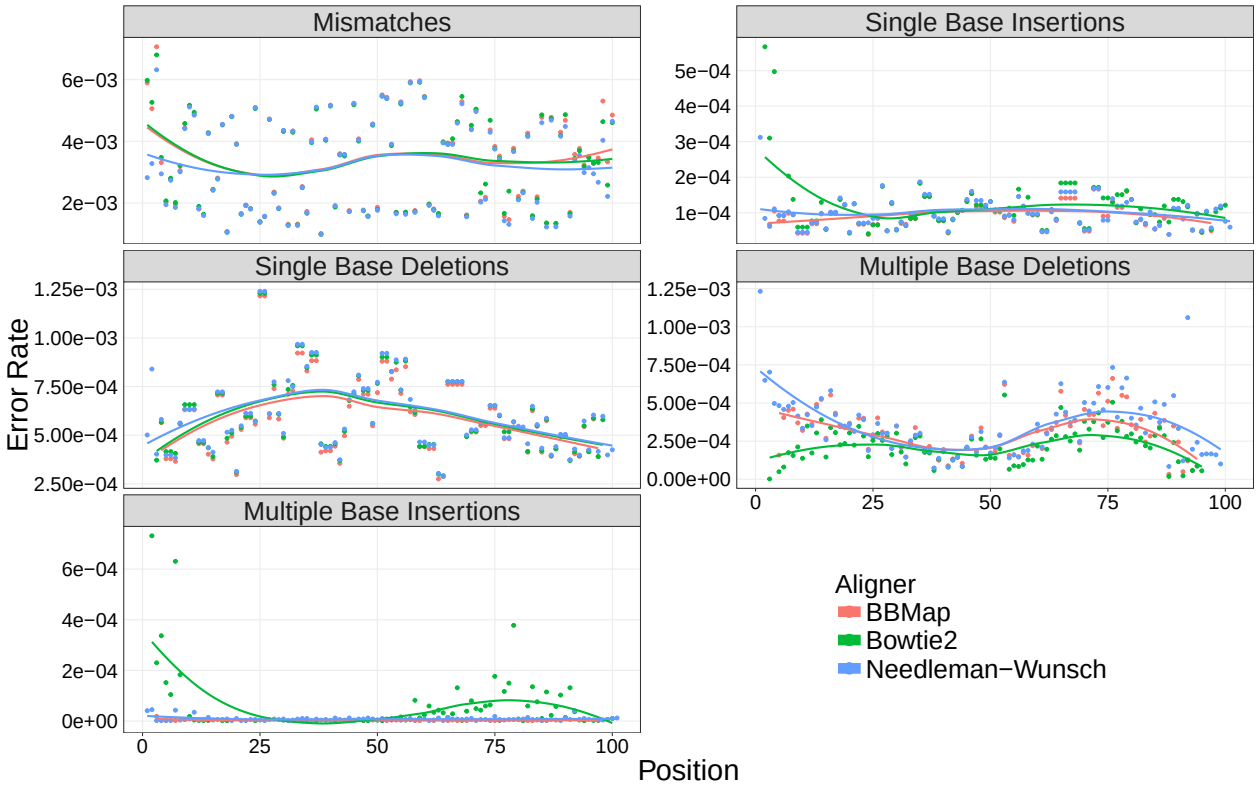

Figure 2.6: **Effect of read aligner on error rates.** Here we mapped reads from the standard IDT oligo with `BBMap` (red), `Bowtie2` (green), and our Needleman-Wunsch aligner (blue), and quantified the error rates with our pipeline. We see that the choice of aligner affects the resulting error rates, especially for detecting multiple-base deletions.

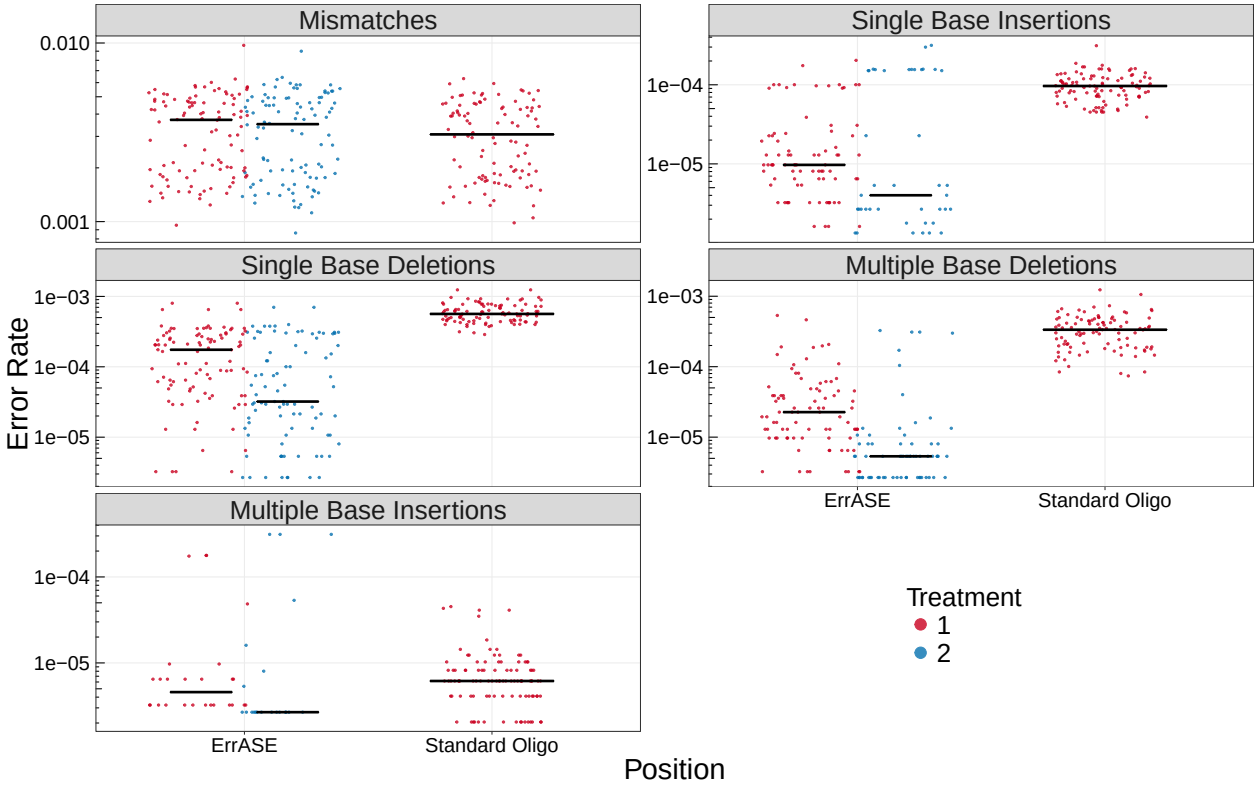

Figure 2.7: **Distributions of error rates per position for the standard oligo assembly before and after ErrASE treatment.** We were unable to detect a significant change between the median error rate after two treatments for mismatches. **Note:** black bar is median value.

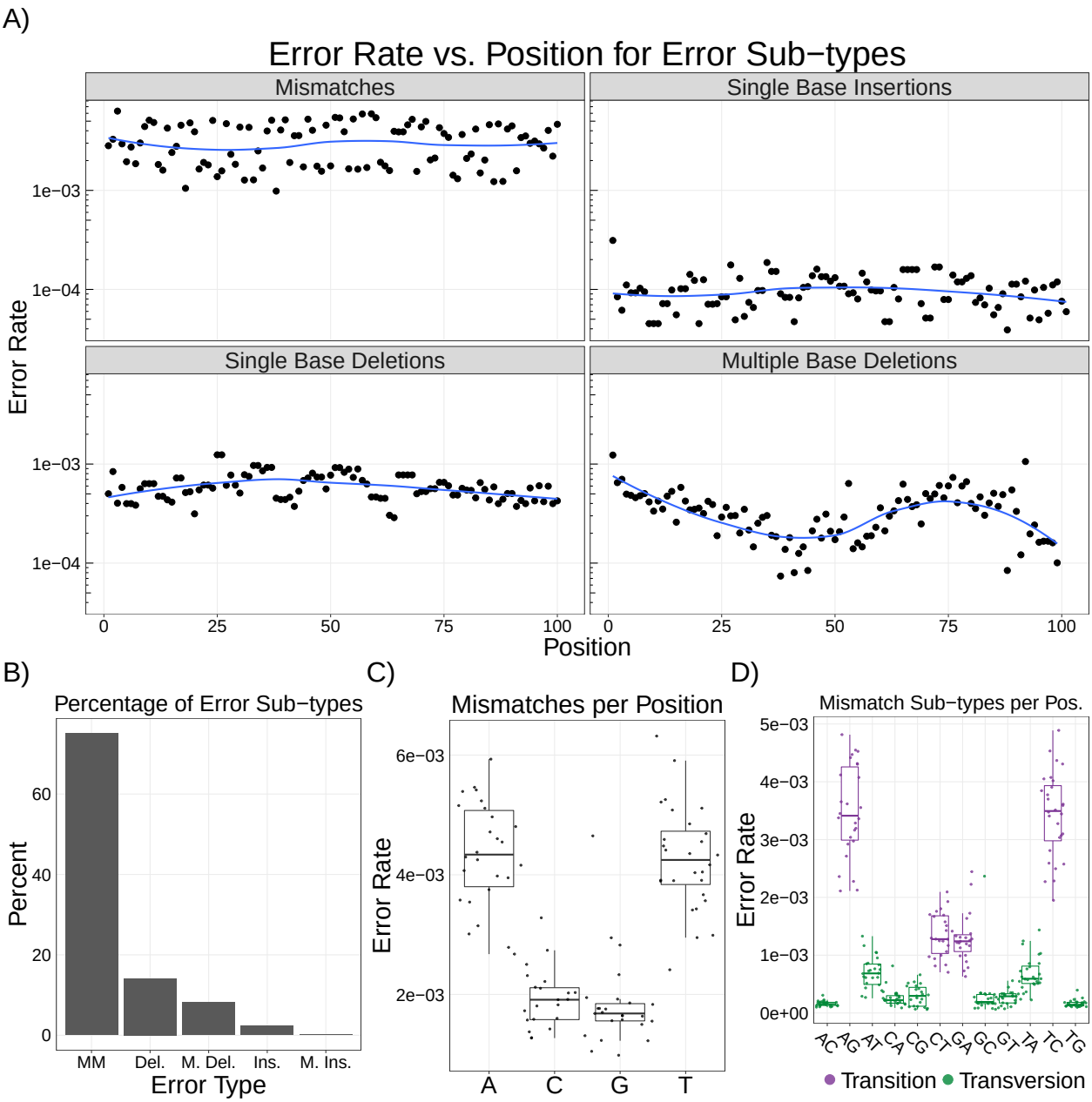

Figure 2.8: **In-depth analysis of standard assemblies. A)** The error rates per base are plotted across each position in our model separated by the four major classes of error types. We do not see strong positional effects for errors across the template. **B)** We find a majority of errors on the template are mismatches (MM), followed by single (Del.) and multiple base (M. Del.) deletions; Single (Ins.) and multiple base (M. Ins.) insertions occur at even lower frequencies. (C) We measure a significantly higher mismatch rate at A's (4.33×10^{-3}) and T's (4.25×10^{-3}) than at G's (1.68×10^{-3}) and C's (1.91×10^{-3}) (Mann-Whitney U, $p << 0.001$). (D) We measure a significantly higher number of transitions (purple) than transversions (green) at each base (Mann-Whitney U, $p << 0.001$). The higher error rates at A's and T's is consistent with Taq polymerase errors. Note: Blue line is a LOESS fit; box plots are first and third quartile for hinges, median for bar, and 1.5× the inter-quartile range for whiskers. **Note:** here we performed the same analysis as Figure 2 in the main text with the error-doped assembly.

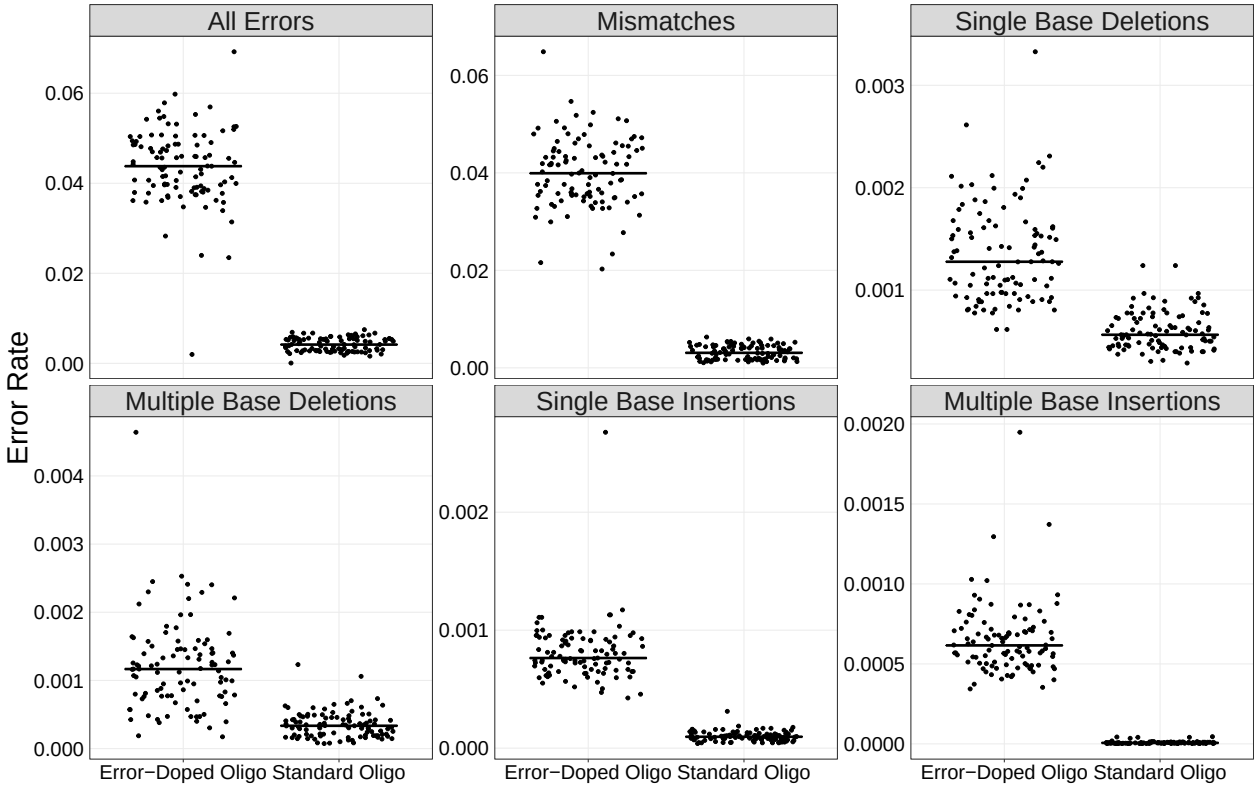

Figure 2.9: **Comparison of measured error rates from error-doped and standard oligos.** Here we plot the distribution of error rates per position and see that for every error sub-type the error rates are significantly higher for the error-doped oligos than those produced by the standard process (Mann-Whitney U Test, all $p << 0.001$). **Note:** Black bar is the median value.

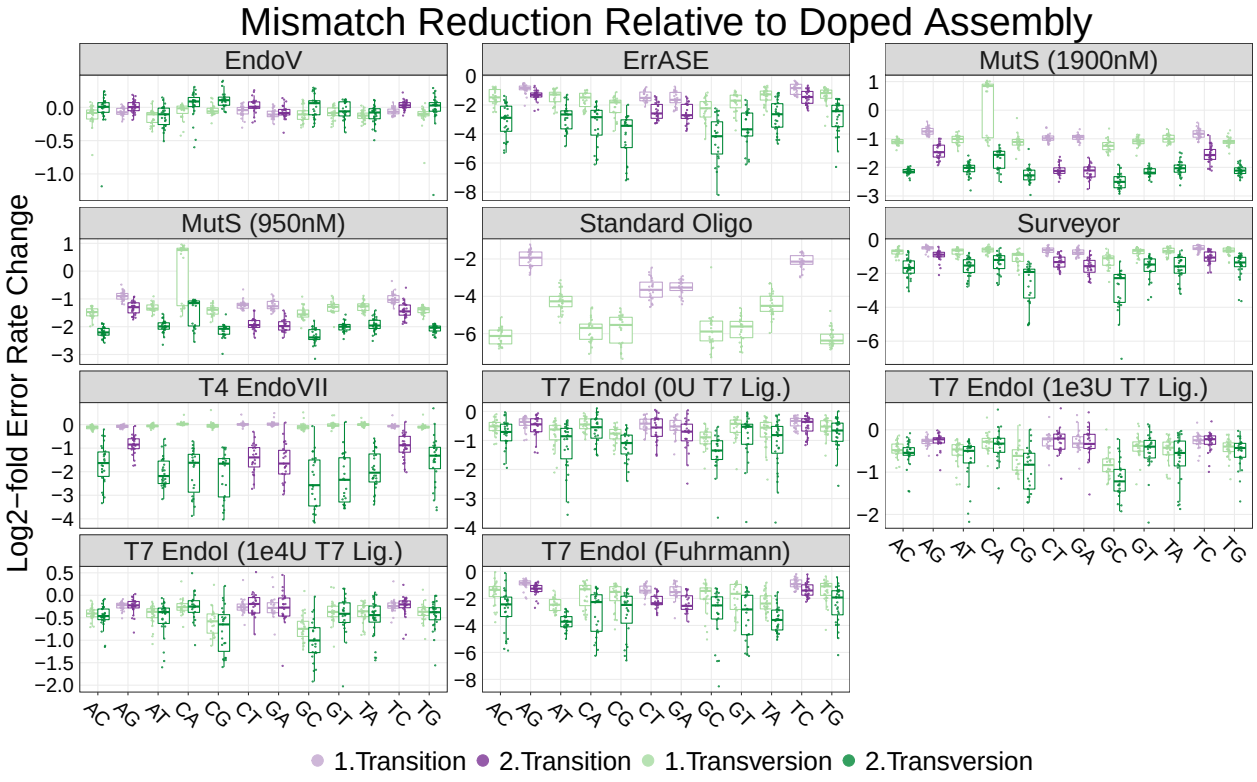

Figure 2.10: **Mismatch correction preferences relative to the error-doped oligo for every enzyme across two consecutive treatments.** Error rates are plotted as the $\log_2$-fold-change in error rate relative to the error-doped template. **Note:** box plots are first and third quartile for hinges, median for bar, and 1.5× the inter-quartile range for whiskers.

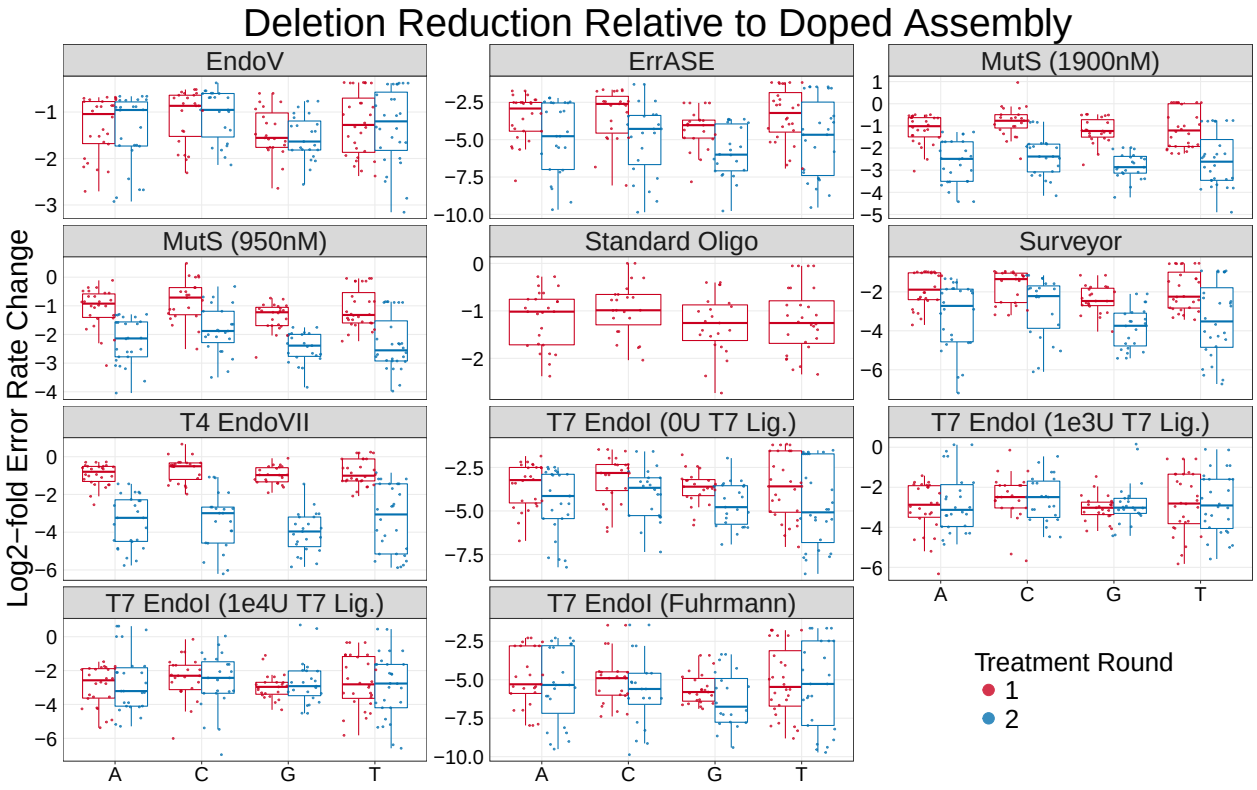

Figure 2.11: **Single-base deletion correction preferences relative to the error-doped oligo for every enzyme across two consecutive treatments.** Error rates are plotted as the $\log_2$-fold-change in error rate relative to the error-doped template. **Note:** box plots are first and third quartile for hinges, median for bar, and 1.5× the inter-quartile range for whiskers.

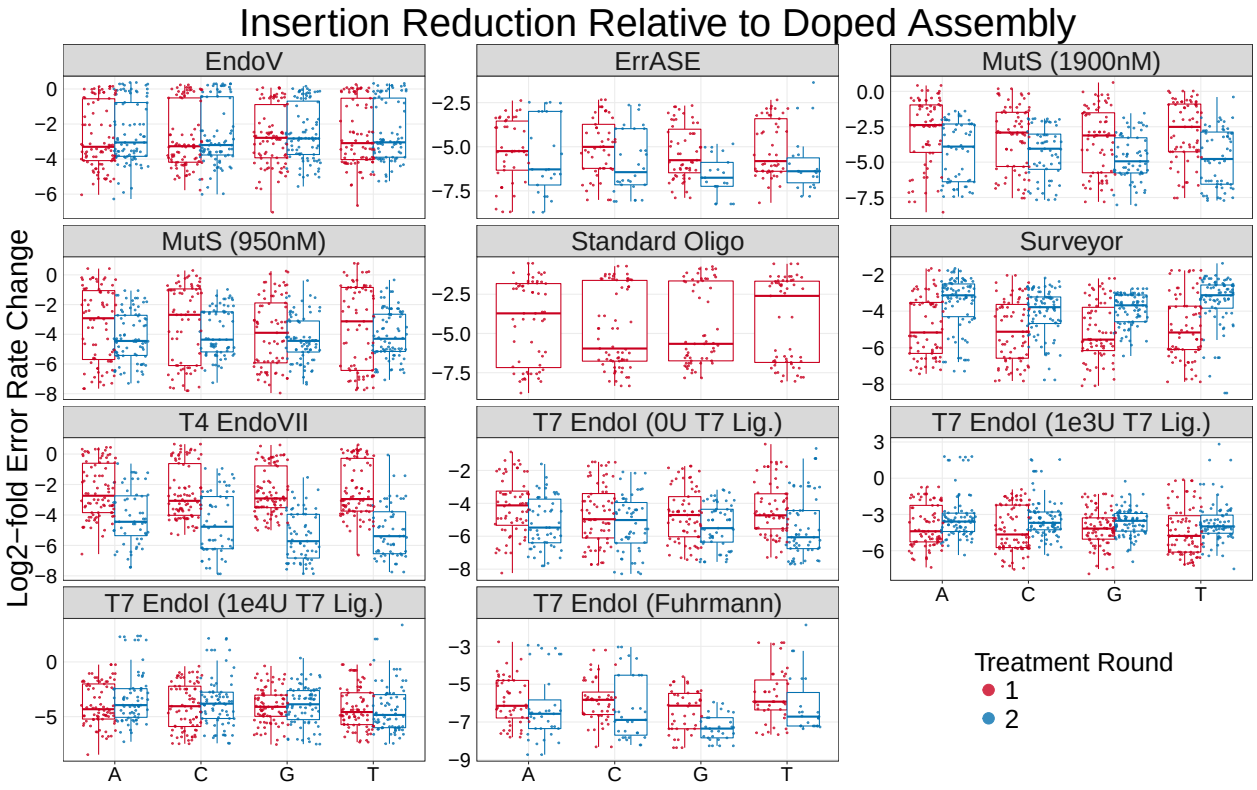

Figure 2.12: **Single-base insertion correction preferences relative to the error-doped oligo for every enzyme across two consecutive treatments.** Error rates are plotted as the $\log_2$-fold-change in error rate relative to the error-doped template. **Note:** box plots are first and third quartile for hinges, median for bar, and 1.5× the inter-quartile range for whiskers.

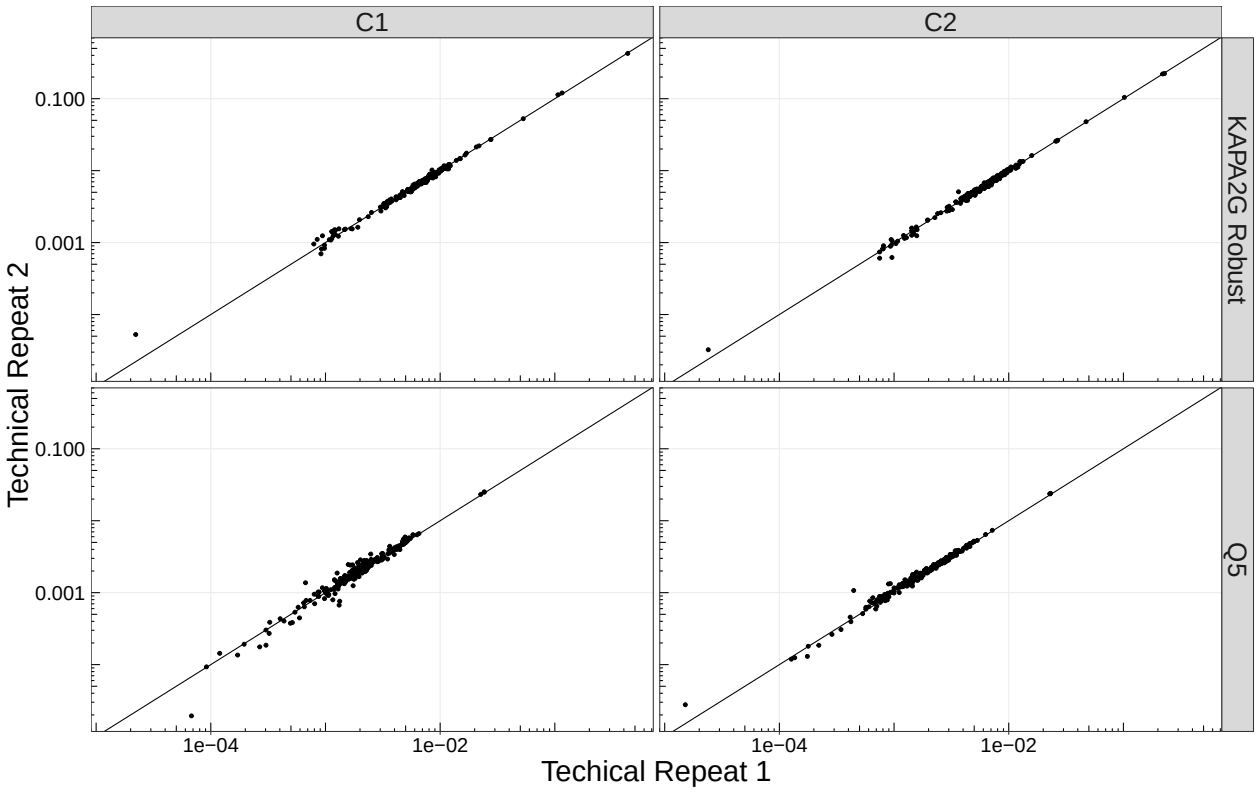

Figure 2.13: **Correlations between error rates for five-oligo assembly technical replicates.** We see that technical replicates are almost perfectly correlated (all $r > 0.995$), with the black line being $y = x$.

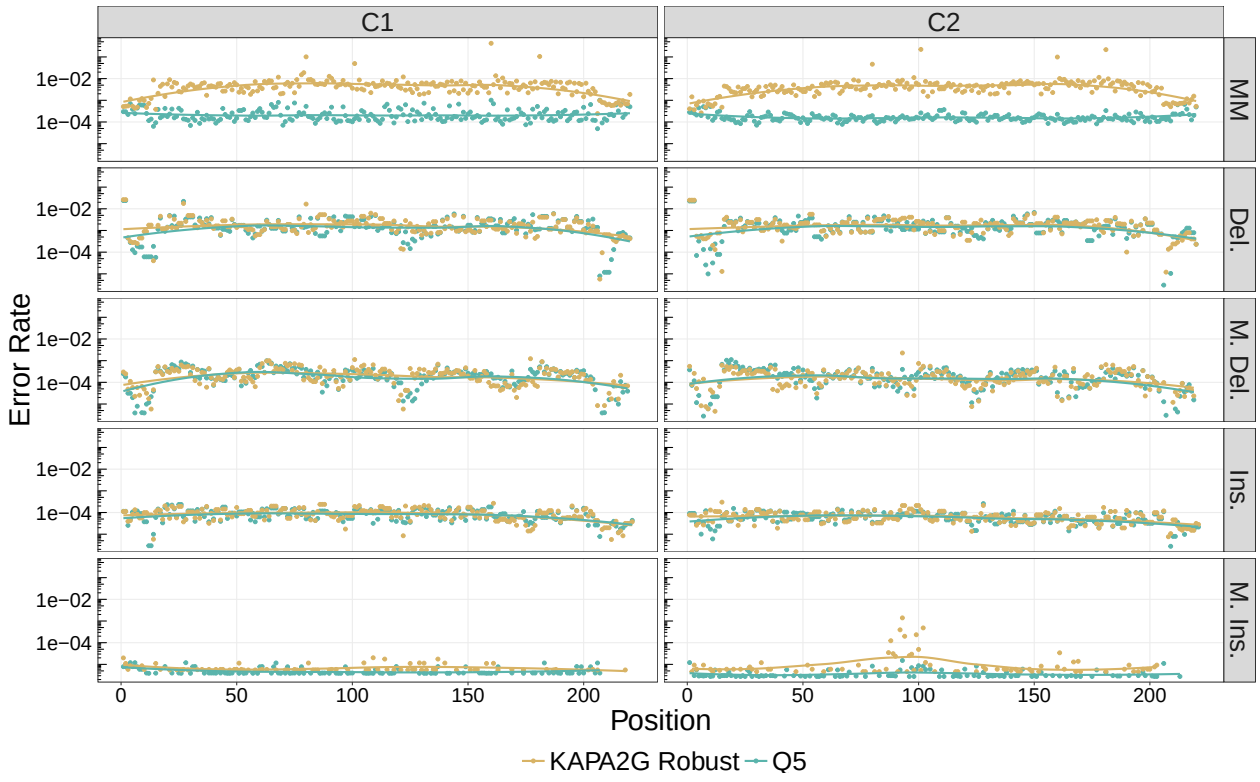

Figure 2.14: **Positional error rate distributions two assemblies using KAPA2G Robust and Q5 polymerase.** We see that KAPA2G Robust, a Taq-based low-fidelity polymerase, incorporates Mismatches (MM) at nearly two-orders of magnitude higher than Q5, a high-fidelity polymerase. We find that both polymerases incorporate single base deletions (Del.), multiple base deletions (M. Del.), single base insertions (Ins.), and multiple base insertions (M. Ins.) at nearly identical rates. With the exception of multiple base insertions, these trends are robust to the different sequence contexts of the two constructs. We note that KAPA2G Robust incorporates a higher number of multiple base insertions around three tandem GGA repeats, likely due to polymerase slippage.

Table 2.2: **Examples of where various aligners fail.** Here _ are padding for visualization, * are soft-trimming, and lower-case bases are inserts.

Aligner:	Ideal	Needleman-Wunsch	Bowtie2	BBMap
Reference:	_GCTGCCGATTT...	_GCTGCCGATTT...	G_CTGCCGATTT...	*GCTGCCGATTT...
Read:	aGCTGCCGATTT...	aGCTGCCGATTT...	aGCTGCCGATTT...	*GCTGCCGATTT...
Reference:	__GCTGCCGATTT...	__GCTGCCGATTT...	G__CTGCCGATTT...	**GCTGCCGATTT...
Read:	aaGCTGCCGATTT...	aaGCTGCCGATTT...	aaGCTGCCGATTT...	**GCTGCCGATTT...
Reference:	GCTGCCGATTT...	GCTGCCGATTT...	GCTGCCGATTT...	GCTGCCGATTT...
Read:	GCT___GATTT...	GCT___GATTT...	___GCTGATTT...	___GCTGATTT...
Reference:	GCTGCCGATTT...	GCTGCCGATTT...	GCTGCCGAT_TT...	GCTGCCGATTT...
Read:	GCTG_____TT...	GCTG_____TT...	______..T...	GCTG_____TT...
Reference:	...TGTATATATCG_	...TGTATATATCG_	...TGTATATATC_G	...TGTATATATCG*
Read:	...TGTATATATCGa	...TGTATATATCGa	...TGTATATATCaG	...TGTATATATCG*
Reference:	...TGTATATATC__G	...TGTATATATC__G	...TGTATATATC__G	...TGTATATATCG**
Read:	...TGTATATATCatG	...TGTATATATCatG	...TGTATATATCatG	...TGTATATATCa**
Reference:	...TGTATATAT__CG	...TGTATATA__TCG	...TGTATATA__TCG	...TGTATATATCG**
Read:	...TGTATATATgtCG	...TGTATATATgtCG	...TGTATATATgtCG	...TGTATATATgt**
Reference:	...TGTATATATCG	...TGTATATATCG	...TGTATATATCG	...TGTATATATCG
Read:	...TGTATATA__G	...TGTATATA__G	...TGTATATAG__	...TGTATATAG__
Reference:	...TGTATATAT__CG	...TGTATATA__TCG	...TGTATATA__TCG	...TGTATATATCG**
Read:	...TGTATATATgtCG	...TGTATATATgtCG	...TGTATATATgtCG	...TGTATATATgt**

Table 2.3: **Median error rate per position for assemblies using the error-doped oligos or the standard oligos.** We measure significant (Mann-Whitney U, $p << 0.001$) differences between the median error rates of the error-doped and standard oligos for all error sub-types.

Type	Error-Doped Oligo	Standard Oligo
All Errors	4.38×10^{-2}	4.18×10^{-3}
Mismatches	3.99×10^{-2}	3.08×10^{-3}
Single Base Deletions	1.28×10^{-3}	5.64×10^{-4}
Multiple Base Deletions	1.17×10^{-3}	3.35×10^{-4}
Single Base Insertions	7.64×10^{-4}	9.65×10^{-5}
Multiple Base Insertions	6.16×10^{-4}	6.16×10^{-6}

References

[1] R. A. Hughes and A. D. Ellington, "Synthetic dna synthesis and assembly: Putting the synthetic in synthetic biology," *Cold Spring Harbor Perspectives in Biology*, vol. 9, no. 1, 2017.

[2] M. Nirenberg and P. Leder, "Rna codewords and protein synthesis," *Science*, vol. 145, no. 3639, pp. 1399–1407, 1964.

[3] S. Kosuri and G. M. Church, "Large-scale de novo DNA synthesis: technologies and applications," *Nature methods*, vol. 11, no. 5, pp. 499–507, 2014.

[4] J. D. Boeke, G. Church, A. Hessel, N. J. Kelley, A. Arkin, Y. Cai, R. Carlson, A. Chakravarti, V. W. Cornish, L. Holt, F. J. Isaacs, T. Kuiken, M. Lajoie, *et al.*, "The genome project-write.," *Science*, 2016.

[5] S. Ma, N. Tang, and J. Tian, "DNA synthesis, assembly and applications in synthetic biology," *Current opinion in chemical biology*, vol. 16, no. 3, pp. 260–267, 2012.

[6] L.-C. Au, F.-Y. Yang, W.-J. Yang, S.-H. Lo, and C.-F. Kao, "Gene synthesis by a lcr-based approach: High-level production of leptin-l54 using synthetic gene inescherichia coli," *Biochemical and biophysical research communications*, vol. 248, no. 1, pp. 200–203, 1998.

[7] X. Zhou, S. Cai, A. Hong, Q. You, P. Yu, N. Sheng, O. Srivannavit, S. Muranjan, J. M. Rouillard, Y. Xia, X. Zhang, Q. Xiang, R. Ganesh, Q. Zhu, A. Matejko, E. Gulari, and X. Gao, "Microfluidic picoarray synthesis of oligodeoxynucleotides and simultaneous assembling of multiple dna sequences," *Nucleic Acids Research*, vol. 32, no. 18, pp. 5409–5417, 2004.

[8] D. Bang and G. M. Church, "Gene synthesis by circular assembly amplification," *Nature methods*, vol. 5, no. 1, pp. 37–39, 2008.

[9] D. G. Gibson, L. Young, R.-Y. Chuang, J. C. Venter, C. A. Hutchison, and H. O. Smith, "Enzymatic assembly of dna molecules up to several hundred kilobases," *Nature methods*, vol. 6, no. 5, pp. 343–345, 2009.

[10] S. L. Beaucage and M. H. Caruthers, "Deoxynucleoside phosphoramidites—a new class of key intermediates for deoxypolynucleotide synthesis," *Tetrahedron Letters*, vol. 22, no. 20, pp. 1859–1862, 1981.

[11] P. A. Carr, J. S. Park, Y.-J. Lee, T. Yu, S. Zhang, and J. M. Jacobson, "Protein-mediated error correction for de novo DNA synthesis," *Nucleic Acids Research*, vol. 32, no. 20, p. e162, 2004.

[12] M. Fuhrmann, W. Oertel, P. Berthold, and P. Hegemann, "Removal of mismatched bases from synthetic genes by enzymatic mismatch cleavage," *Nucleic acids research*, vol. 33, no. 6, pp. e58–e58, 2005.

[13] A. Currin, N. Swainston, P. J. Day, and D. B. Kell, "Speedygenes: an improved gene synthesis method for the efficient production of error-corrected, synthetic protein libraries for directed evolution," *Protein Engineering Design and Selection*, vol. 27, no. 9, pp. 273–280, 2014.

[14] A. F. Sequeira, C. I. Guerreiro, R. Vincentelli, and C. M. Fontes, "T7 endonuclease i mediates error correction in artificial gene synthesis," *Molecular Biotechnology*, vol. 58, no. 8-9, pp. 573–584, 2016.

[15] S. Kosuri, N. Eroshenko, E. M. LeProust, M. Super, J. Way, J. B. Li, and G. M. Church, "Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips," *Nature biotechnology*, vol. 28, no. 12, pp. 1295–1299, 2010.

[16] J. Tian, H. Gong, N. Sheng, X. Zhou, E. Gulari, X. Gao, and G. Church, "Accurate multiplex gene synthesis from programmable DNA microchips," *Nature*, vol. 432, no. 7020, pp. 1050–1054, 2004.

[17] J. Quan, I. Saaem, N. Tang, S. Ma, N. Negre, H. Gong, K. P. White, and J. Tian, "Parallel on-chip gene synthesis and application to optimization of protein expression," *Nature biotechnology*, vol. 29, no. 5, pp. 449–452, 2011.

[18] H. Kim, H. Han, J. Ahn, J. Lee, N. Cho, H. Jang, H. Kim, S. Kwon, and D. Bang, "'shotgun DNA synthesis' for the high-throughput construction of large DNA molecules," *Nucleic acids research*, p. gks546, 2012.

[19] I. Saaem, S. Ma, J. Quan, and J. Tian, "Error correction of microchip synthesized genes using surveyor nuclease," *Nucleic Acids Research*, vol. 40, no. 3, p. e23, 2012.

[20] W. Wan, L. LI, Q. Xu, Z. Wang, Y. Yao, R. Wang, J. Zhang, H. Liu, X. Gao, and J. Hong, "Error removal in microchip-synthesized DNA using immobilized muts," *Nucleic Acids Research*, vol. 42, no. 12, p. e102, 2014.

[21] A. Ellington and J. D. Pollard, "Introduction to the synthesis and purification of oligonucleotides," *Current Protocols in Nucleic Acid Chemistry*, pp. A–3C, 2001.

[22] N. D. Sinha and K. E. Jung, "Analysis and purification of synthetic nucleic acids using hplc," *Current Protocols in Nucleic Acid Chemistry*, pp. 10–5, 2015.

[23] A. Y. Borovkov, A. V. Loskutov, M. D. Robida, K. M. Day, J. A. Cano, T. Le Olson, H. Patel, K. Brown, P. D. Hunter, and K. F. Sykes, "High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides," *Nucleic Acids Research*, vol. 38, no. 19, p. e180, 2010.

[24] M. Matzas, P. F. Stähler, N. Kefer, N. Siebelt, V. Boisguérin, J. T. Leonard, A. Keller, C. F. Stähler, P. Häberle, B. Gharizadeh, *et al.*, "High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing," *Nature biotechnology*, vol. 28, no. 12, pp. 1291–1294, 2010.

[25] J. J. Schwartz, C. Lee, and J. Shendure, "Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules," *Nature methods*, vol. 9, no. 9, pp. 913–915, 2012.

[26] H. Lee, H. Kim, S. Kim, T. Ryu, H. Kim, D. Bang, and S. Kwon, "A high-throughput optomechanical retrieval method for sequence-verified clonal dna from the ngs platform," *Nature communications*, vol. 6, p. 6073, 2015.

[27] R. D. Mashal, J. Koontz, and J. Sklar, "Detection of mutations by cleavage of dna heteroduplexes with bacteriophage resolvases.," *Nature genetics*, vol. 9, no. 2, pp. 177–183, 1995.

[28] R. Youil, B. W. Kemper, and R. Cotton, "Screening for mutations by enzyme mismatch cleavage with t4 endonuclease vii," *Proceedings of the National Academy of Sciences*, vol. 92, no. 1, pp. 87–91, 1995.

[29] R. Wagner, P. Debble, and M. Radman, "Mutation detection using immobilized mismatch binding protein (muts)," *Nucleic acids research*, vol. 23, no. 19, pp. 3944–3948, 1995.

[30] P. Qiu, H. Shandilya, J. M. D Alessio, K. O Connor, J. Durocher, and G. F. Gerard, "Mutation detection using surveyor nuclease," *Biotechniques*, vol. 36, no. 4, pp. 702–707, 2004.

[31] J. Smith and P. Modrich, "Removal of polymerase-produced mutant sequences from pcr products," *Proceedings of the National Academy of Sciences*, vol. 94, no. 13, pp. 6847–6850, 1997.

[32] A. Whitehouse, J. Deeble, R. Parmar, G. R. Taylor, A. F. Markham, and D. M. Meredith, "Analysis of the mismatch and insertion/deletion binding properties of thermus thermophilus, hb8, muts," *Biochemical and Biophysical Research Communications*, vol. 233, no. 3, pp. 834–837, 1997.

[33] J. Brown, T. Brown, and K. R. Fox, "Affinity of mismatch-binding protein muts for heteroduplexes containing different mismatches," *Biochemical Journal*, vol. 354, no. 3, pp. 627–633, 2001.

[34] M. Cho, S. Chung, S.-D. Heo, J. Ku, and C. Ban, "A simple fluorescent method for detecting mismatched {DNAs} using a muts–fluorophore conjugate," *Biosensors and Bioelectronics*, vol. 22, no. 7, pp. 1376–1381, 2007.

[35] F. S. Groothuizen, A. Fish, M. V. Petoukhov, A. Reumer, L. Manelyte, H. H. K. Winterwerp, M. G. Marinus, J. H. G. Lebbink, D. I. Svergun, P. Friedhoff, and T. K. Sixma, "Using stable muts dimers and tetramers to quantitatively analyze DNA mismatch recognition and sliding clamp formation," *Nucleic Acids Research*, vol. 41, no. 17, pp. 8166–8181, 2013.

[36] D. H. Geschwind, R. Rhee, and S. F. Nelson, "A biotinylated muts fusion protein and its use in a rapid mutation screening technique," *Genetic analysis: biomolecular engineering*, vol. 13, no. 4, pp. 105–111, 1996.

[37] S. Ma, I. Saaem, and J. Tian, "Error correction in gene synthesis technology," *Trends in biotechnology*, vol. 30, no. 3, pp. 147–154, 2012.

[38] T. Tsuji and Y. Niida, "Development of a simple and highly sensitive mutation screening system by enzyme mismatch cleavage with optimized conditions for standard laboratories," *ELECTROPHORESIS*, vol. 29, pp. 1473–1483, Apr. 2008.

[39] M. C. Huang, W. C. Cheong, L. S. Lim, and M.-H. Li, "A simple, high sensitivity mutation screening using ampligase mediated t7 endonuclease i and surveyor nuclease with microfluidic capillary electrophoresis," *ELECTROPHORESIS*, vol. 33, no. 5, pp. 788–796, 2012.

[40] L. Vouillot, A. Thélie, and N. Pollet, "Comparison of t7e1 and surveyor mismatch cleavage assays to detect mutations triggered by engineered nucleases," *G3: Genes|Genomes|Genetics*, 2015.

[41] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.

[42] S. F. Altschul and B. W. Erickson, "Optimal sequence alignment using affine gap costs," *Bulletin of Mathematical Biology*, vol. 48, no. 5, pp. 603–616, 1986.

[43] H. Li and R. Durbin, "Fast and accurate long-read alignment with burrows-wheeler transform," *Bioinformatics*, vol. 26, no. 5, pp. 589–595, 2010.

[44] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature methods*, vol. 9, no. 4, pp. 357–359, 2012.

[45] P. Keohavong and W. G. Thilly, "Fidelity of dna polymerases in dna amplification," *Proceedings of the National Academy of Sciences*, vol. 86, no. 23, pp. 9253–9257, 1989.

[46] M. S. Hestand, J. V. Houdt, F. Cristofoli, and J. R. Vermeesch, "Polymerase specific error rates and profiles identified by single molecule sequencing," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 784–785, pp. 39–45, 2016.

[47] P. McInerney, P. Adams, and M. Z. Hadi, "Error rate comparison during polymerase chain reaction by dna polymerase," *Molecular biology international*, vol. 2014, 2014.

[48] V. Potapov and J. L. Ong, "Examining sources of error in pcr by single-molecule sequencing," *PLOS ONE*, vol. 12, pp. 1–19, 01 2017.

[49] R. Saiki, D. Gelfand, S. Stoffel, S. Scharf, R. Higuchi, G. Horn, K. Mullis, and H. Ehrlich, "Primer-directed enzymatic amplification of dna," *Science*, vol. 239, pp. 487–491, 1988.

[50] D. F. Lee, J. Lu, S. Chang, J. J. Loparo, and X. S. Xie, "Mapping DNA polymerase errors by single-molecule sequencing," *Nucleic Acids Research*, vol. 44, no. 13, p. e118, 2016.

[51] S. Carson, S. T. Wick, P. A. Carr, M. Wanunu, and C. A. Aguilar, "Direct analysis of gene synthesis reactions using solid-state nanopores," *ACS nano*, vol. 9, no. 12, pp. 12417–12424, 2015.

[52] A. Chakraborty and S. Bandyopadhyay, "Fogsaa: Fast optimal global sequence alignment algorithm," *Scientific reports*, vol. 3, p. 1746, 2013.

[53] B. Bushnell, "BBMap:BBMap short read aligner, and other bioinformatic tools." `https://sourceforge.net/projects/bbmap/`.

[54] R. Hart, "uta-align provides C-based Needleman-Wunsch and Smith-Waterman alignment algorithms with a Python interface.." `https://github.com/biocommons/uta-align`.

[55] H. Wickham, "Tidy data," *Journal of Statistical Software*, vol. 59, no. 1, pp. 1–23, 2014.

[56] N. Eroshenko, S. Kosuri, A. H. Marblestone, N. Conway, and G. M. Church, "Gene assembly from chip-synthesized oligonucleotides," *Current Protocols in Chemical Biology*, 2009.

[57] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[58] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

Chapter 3

Multiplexed Gene Synthesis in Emulsions for Exploring Protein Functional Landscapes

3.1 Abstract

Improving our ability to construct and functionally characterize DNA sequences would broadly accelerate progress in biology. Here, we introduce DropSynth, a scalable, low-cost method to build thousands of defined gene-length constructs in a pooled (multiplexed) manner. DropSynth uses a library of barcoded beads that pull down the oligonucleotides necessary for a gene's assembly, which are then processed and assembled in water-in-oil emulsions. We use DropSynth to successfully build >7000 synthetic genes that encode phylogenetically-diverse homologs of two essential genes in *E. coli*. We tested the ability of phosphopantetheine adenylyltransferase homologs to complement a knockout *E. coli* strain in multiplex, revealing core functional motifs and reasons underlying homolog incompatibility. DropSynth coupled with multiplexed functional assays allow us to rationally explore sequence-function relationships at unprecedented scale.

This chapter has been published as: C. Plesa[†], **A. M. Sidore**[†], N. B. Lubock, D. Zhang, and S. Kosuri "Multiplexed gene synthesis in emulsions for exploring protein functional landscapes," *Science*, vol. 359, no. 6373, pp. 343-347, 2018

3.2 Main Text

The scale at which we can build and functionally characterize DNA sequences sets the pace at which we explore and engineer biology. The recent development of multiplexed functional assays allows for the facile testing of thousands to millions of sequences across a wide array of biological functions [1, 2]. Currently, such assays are limited by their ability to build or access DNA sequences to test. Natural or mutagenized DNA sequences [3, 4] allow for large libraries, but are not easily programmed and thus limit hypotheses, applications, and engineered designs. Alternatively, researchers can use low-cost microarray-based oligo pools that allow for large libraries of designed ∼200 nucleotide (nt) sequences [5], but their short lengths limit many other applications. Gene synthesis is capable of creating long-length sequences, but high costs currently prohibit building large libraries of designed sequences [6, 7, 8, 9].

Here we develop a gene synthesis method we term DropSynth, a multiplexed approach capable of building large pooled libraries of designed gene-length sequences. DropSynth uses microarray derived oligo libraries to assemble gene libraries at vastly reduced costs. We and others have developed robust parallel processes to build genes from oligo arrays, but because each gene must be assembled individually, costs are prohibitive for large gene libraries [6, 10]. In these efforts, the ability to isolate and concentrate DNA from the background pool complexity was paramount for robust assemblies [11]. Previous efforts to multiplex such assemblies have not isolated reactions from one another, and thus suffered from short assembly lengths, highly-biased libraries, the inability to scale, and constraints on sequence homology [12, 13, 14, 15].

DropSynth works by pulling down only those oligos required for a particular gene's assembly onto barcoded microbeads from a complex oligo pool. By emulsifying this mixture into picoliter droplets, we isolate and concentrate the oligos prior to gene assembly, thereby overcoming the critical roadblocks for proper assembly and scalability (Fig. 3.1A, Supplemental Movie S1). The microbead barcodes are unique 12 nt sequences that all oligos for a particular assembly share, and pair with complementary strands displayed on the microbead. Within each droplet, sequences are released from the bead using Type IIs restriction enzyme sites and assembled through polymerase cycling assembly (PCA) into full length genes. Finally, the emulsion is broken and the gene library is recovered. To test and optimize the protocol, we built model assemblies that were unique, but

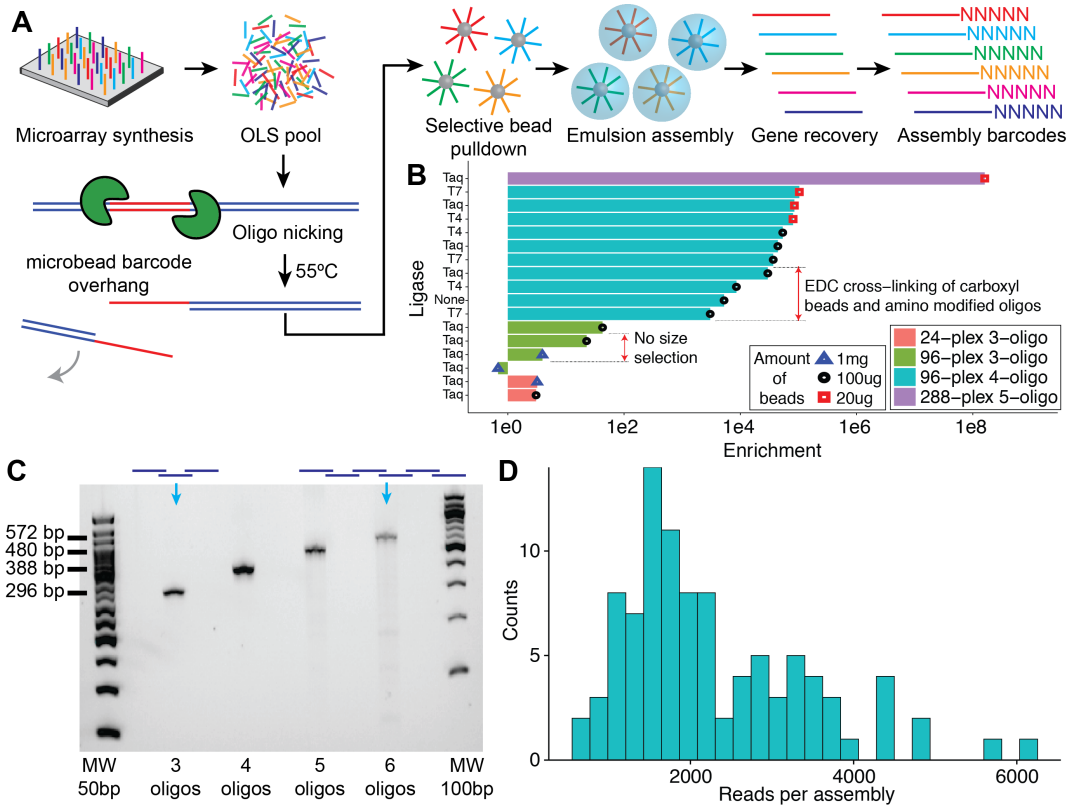

Figure 3.1: **DropSynth assembly and optimization. A.** We amplified array-derived oligos and exposed a single-stranded region that acts as a gene-specific microbead barcode. Barcoded beads display complementary single-stranded regions that selectively pull down the oligos necessary to assemble each gene. The beads are then emulsified, and the oligos are assembled by PCA. The emulsion is then broken, and the resultant assembled genes are barcoded and cloned. **B.** We used a model gene library that allowed us to monitor the level of specificity and coverage of the assembly process. We then optimized various aspects of the protocol including purification steps, DNA ligase, and bead couplings to improve the specificity of the assembly reaction. Enrichment is defined as the number of specific assemblies observed relative to what would be observed by random chance in a full combinatorial assembly. **C.** We attempted 96-plex gene assemblies with 3, 4, 5, or 6 oligonucleotides and the resultant libraries displayed the correct-sized band on an agarose gel. **D.** The distribution of read-counts for all 96 assemblies (4-oligo assembly) as determined by NGS.

shared common overlap sequences. As a result, any contaminating oligo would still participate in the assembly reaction, allowing us to monitor assembly specificity and library coverage. We optimized each aspect of the protocol by trying to assemble 24-, 96-, and 288-member libraries composed of 3, 4, 5, and 6 oligos at once, based on how often we saw intended targets versus their expected frequency given random (i.e. bulk) assembly (Fig. 3.1B). Over many iterations we achieved high enrichment rates ($\sim 10^8$) by modifying the amount of beads, presence of size selection after assembly, ligase used for capture, and type of bead chemistry, testing both EDC crosslinking of carboxyl beads and streptavidin-coupled beads. We ultimately found that using streptavidin bead chemistry, Taq ligase for bead capture, and size-selection after assembly yielded the highest enrichment rates. Using these protocols, we were able to build libraries of up to 6 oligos that produced correct sized bands (Fig. 3.1C), and the resulting assembly distributions were not overly skewed (Fig. 3.1D, Fig. 3.5).

To test the scalability of DropSynth, we attempted assembly of 12,672 genes ranging in size from 381 to 669 bp which encode homologs of two bacterial proteins from across the tree of life (Fig. 3.2A, Fig. 3.6). A total of 33 libraries of 384 genes each encoded 5,775 homologs of dihydrofolate reductase (DHFR) with two different codon usages (11,520 DHFR genes), as well as 1,152 homologs of the enzyme phosphopantetheine adenylyltransferase (PPAT) (Fig. 3.7, A and B). DHFR genes were assembled from either four or five 230-mer oligos while PPAT genes were assembled from five 200-mer oligos. We obtained correctly-sized bands for 31/33 assemblies, with one failing due to oligo amplification issues and the other due to low yield on the oligo processing steps, in contrast to attempts using bulk assembly which produced shorter failed by-products (Fig. 3.7C). Three of the libraries (5x 230-mers) were too long to verify using our barcoding approach, but the resulting synthesis showed correct band formation (Fig. 3.8).

We cloned the libraries into an expression plasmid containing a random 20 bp barcode (assembly barcode) and sequenced the remaining 28 libraries consisting of 10,752 designs (Fig. 3.7D and Fig. 3.8, Fig. 3.9). For the PPAT 5x 200-mer assemblies, sequencing revealed that a total of 872 genes (75%) had assemblies corresponding to a perfect amino acid sequence represented by at least one assembly barcode, with a median of 2 reads per assembly barcode and 56 assembly barcodes per homolog (Fig. 3.2B, Fig. 3.10, A and B). This coverage increased when including sequences with deviations from the designed sequences, with 1,002 genes (87%) represented within 5 aa from the

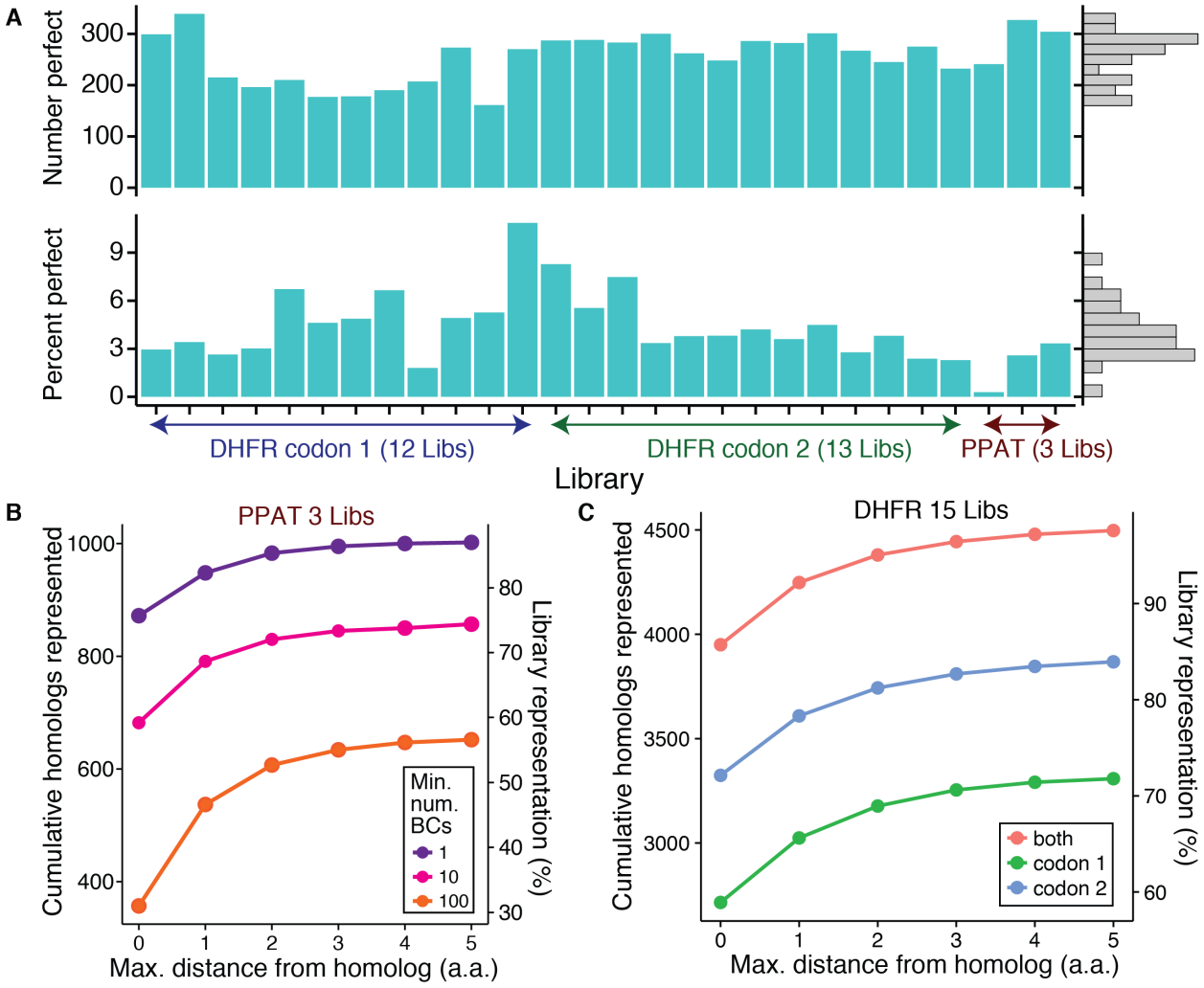

Figure 3.2: **DropSynth assembly of 10,752 genes. A.** We used DropSynth to assemble 28 libraries of 10,752 genes representing 1,152 homologs of PPAT and 4,992 homologs of DHFR. The number of library members with at least one perfect assembly and the median percent perfects determined using constructs with at least 100 barcodes is shown for each library. **B.** We observe that 872 PPAT homologs (75%) had at least one perfect assembly, and 1,002 homologs (87%) had at least one assembly within a distance of 5 a.a. from design. **C.** We assembled two codon variants for each designed DHFR homolog, allowing us to achieve higher coverage.

designed sequences (all homologs have some alignments regardless of distance) (Fig. 3.10D). For the DHFR 4x 230-mer assemblies we observed perfect sequences for 65% (6,271) of the designed homologs, and 75% have at least one assembly within 2 aa difference from design. Since there are two codon usages per homolog, when combined over homologs we observe 3,950 (79%) have at least one perfect, and 88% have at least one assembly in distance 2 aa (Fig. 3.2C). We see a strong correlation (ρ=0.73 (Pearson), p-value=3.4E-5) between the amount of DNA used to load the DropSynth beads and the resulting library coverage (Fig. 3.11A). We also found 15 microbead barcodes that have more dropouts than would be expected by chance (Fig. 3.11B). For constructs with at least 100 assembly barcodes, we observed a median of 1.9% ($\sigma = 2.9\%$) and 3.9% ($\sigma = 3.8\%$) perfect protein assemblies (Fig. 3.2A, Fig. 3.10C, Fig. 3.12) for PPAT and DHFR libraries respectively. The nearly double the rate of perfects for DHFR libraries compared to PPAT can be attributed to using longer oligos (230 vs. 200 nt) that only require 4 oligos instead of 5 to assemble the gene (Fig. 3.13A). Increasing the oligo length provides a way to assemble longer genes without significant decreases in the resulting yields (Fig. 3.13B). Furthermore, the distribution of perfect assemblies in the PPAT libraries is not overly skewed (Fig. 3.10D) and most library members have assemblies with high identity to their respective designed homologs (Fig. 3.10F). The resultant error profiles were consistent with Taq-derived mismatch and assembly errors that we have observed previously [16] (Fig. 3.14).

We sought to show how DropSynth-assembled libraries could be easily coupled as inputs into multiplex functional assays by probing how well the PPAT homologs of various evolutionary distance to *E. coli* could rescue a knockout phenotype. PPAT is an essential enzyme, encoded by the gene *coaD*, which catalyzes the 2$^{\text{nd}}$ to last step in the biosynthesis of coenzyme A (CoA) [17] (Fig. 3.15) and is an attractive target for the development of novel antibiotics [18]. Assembled PPAT variants on the barcoded expression plasmid were transformed into *E. coli* Δ*coaD* cells and screened for complementation by growing the library in batch culture through three serial 1000-fold dilutions (Fig. 3.3A, Table 3.1), while a rescue plasmid was simultaneously heat cured (Fig. 3.16). Assembly barcode sequencing of the resulting populations provided a reproducible estimate for the fitness of all homologs successfully assembled without error (biological replicates ρ=0.94; Pearson, p-value <2.2E-16) (Fig. 3.17A, Fig. 3.18A). Individual barcodes can display considerable noise, so having many assembly barcodes per construct improved confidence (Fig. 3.18, B and C). Negative controls

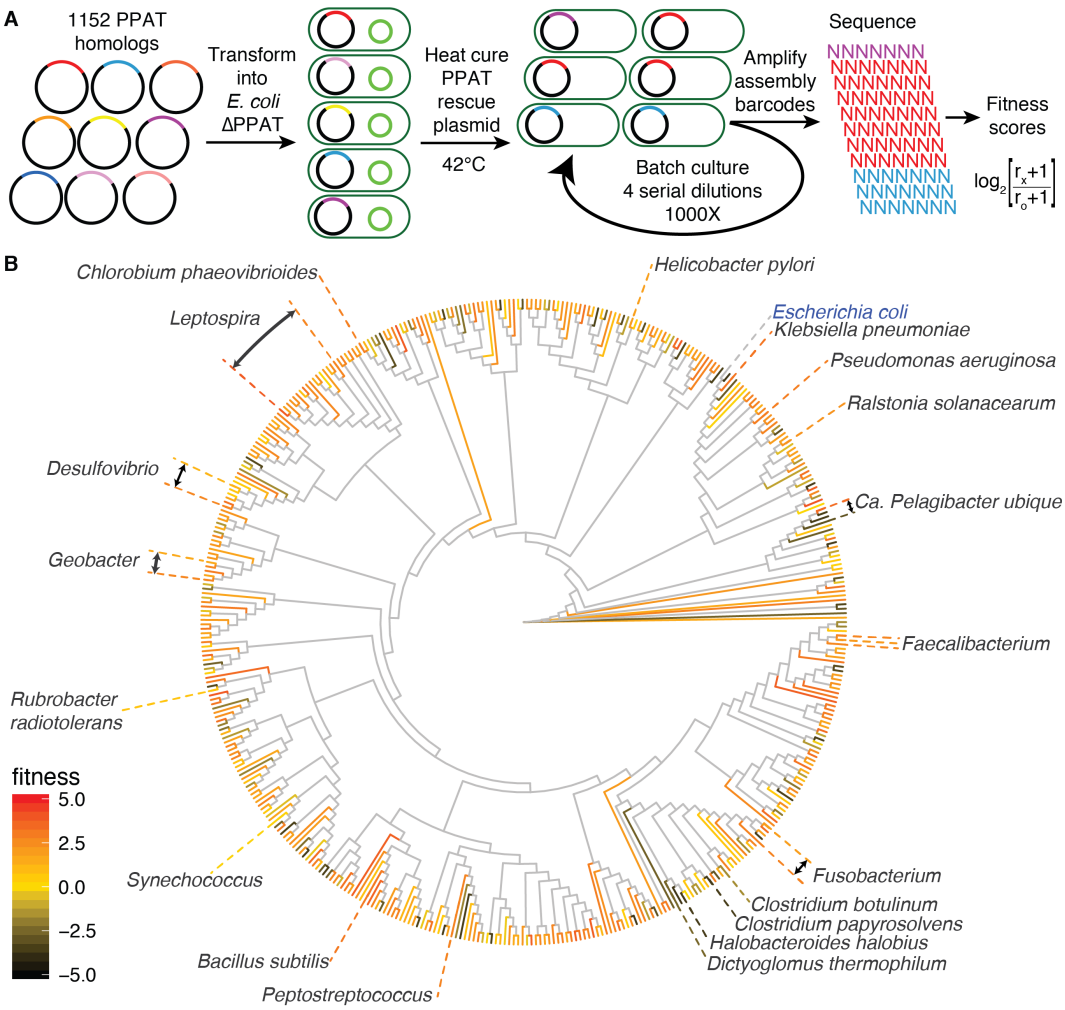

Figure 3.3: **PPAT complementation assay. A.** We used DropSynth to assemble a library of 1152 homologs of phosphopantetheine adenylyltransferase (PPAT), an essential enzyme catalyzing the second-to-last step in coenzyme A biosynthesis, and functionally characterized them using a pooled complementation assay. The barcoded library was transformed into *E. coli* Δ*coaD* cells containing a curable rescue plasmid expressing *E. coli coaD*. The rescue plasmid was removed allowing the homologs and their mutants to compete with each other in a batch culture. We tracked assembly barcode frequencies over four serial 1000-fold dilutions, and used the frequency changes to assign a fitness score. **B.** This phylogenetic tree shows 451 homologs each with at least 5 assembly barcodes, a subset of the full data set, where leaves are colored by fitness. Despite having a median 50% sequence identity, we find that the majority of PPAT homologs are able to complement the function of the native *E. coli* PPAT, with 70% having positive fitness values, while low-fitness homologs are dispersed throughout the tree without much clustering of clades.

and sequences containing indels show strong depletion (Fig. 3.17A, Fig. 3.19A, 3.20), and fitness is reduced with increasing numbers of mutations (ρ=-0.38; Spearman, p-value <2.2E-16) (Fig. 3.19, B and C). Pooled fitness scores also correlated well with measured growth rates of individually tested controls (r_s=0.86, Spearman, p-value 5.9E-12) (Fig. 3.21). Approximately 14% percent of the homologs show strong depletion (fitness below -2.5) while 70% have a positive fitness value in the pooled assay. Low-fitness homologs are evenly distributed throughout the phylogenetic tree with only minor clustering of clades (Fig. 3.3B, Fig. 3.17B, Fig. 3.22, 3.23A) showing the high modularity of PPAT. There are several reasons homologs could have low fitness including environmental mismatches, improper folding, mismatched metabolic flux, interactions with other cytosolic components, or gene dosage toxicity effects resulting from improperly high expression [19] (Supplementary Text).

Errors during the oligo synthesis or DropSynth assembly give us mutational data across all the homologs, which we can further analyze to better understand function. We selected all 497 homologs that showed some degree of complementation (fitness greater than -1) as well as their 71,061 mapped mutants within distance 5 a.a. and carried out a multiple sequence alignment to find equivalent residue positions. For each amino acid and position, we found the median fitness among all of these homologs and mutants. The resulting data was projected onto the *E. coli* PPAT sequence (Fig. 3.4A and B), providing data similar to deep mutational scanning approaches [22, 23]. We term this approach broad mutational scanning (BMS). The average BMS fitness for each position shows strong constraints in the catalytic site, at highly conserved sites (ρ=-0.64; Pearson, p-value <2.2E-16), and at buried residues compared to solvent-accessible ones (ρ=0.42; Pearson, p-value 3.9E-8) (Fig. 3.24, A and B, Supplementary Text). Surprisingly, some residues that are known to interact with either ATP or 4'-phosphopantetheine turn out to be relatively promiscuous when averaged over a large number of homologs. Furthermore, when mapped onto the *E. coli* structure (Fig. 3.4B), positions known to be involved with allosteric regulation by coenzyme A or dimer formation, show relatively little constraint, highlighting the diversity of distinct approaches employed among different homologs, while maintaining the same core function. We implemented a simple binary classifier to predict the sign of the BMS fitness value based on a number of features, achieving an accuracy of 0.825 (Fig. 3.25).

Additionally, we can search for gain-of-function (GoF) mutations amongst those homologs that

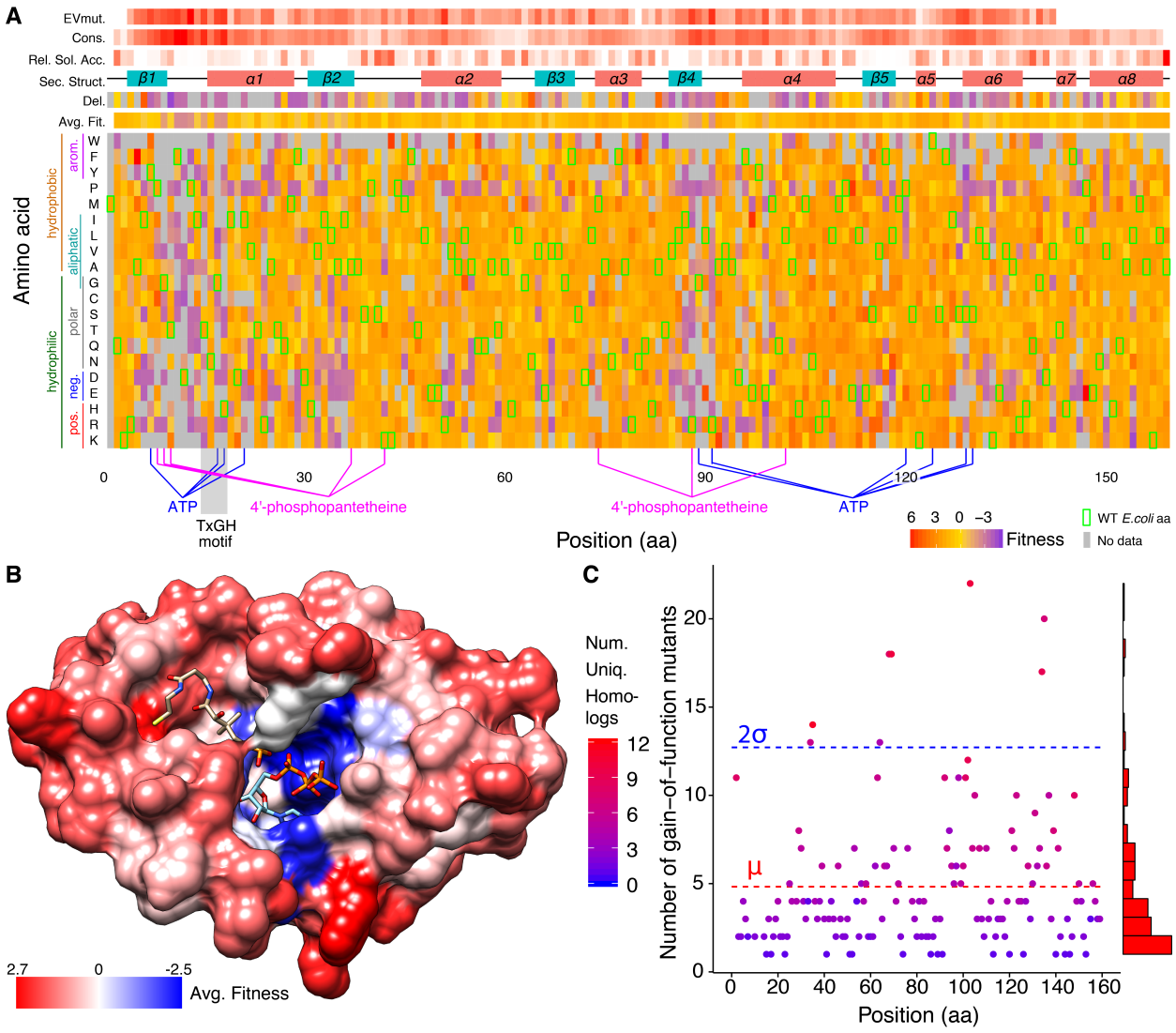

Figure 3.4: **Broad mutational scanning (BMS) analysis. A.** The fitness landscape of 497 complementing PPAT homologs and their 71,061 mutants (within a distance of 5 a.a.) is projected onto the *E. coli* PPAT sequence, with each point in the heatmap showing the average fitness over all sequences containing that amino acid at each aligned position. Mutations are highly constrained at a core group of residues involved in catalytic function. Other positions show relatively little loss of function, when averaged over many homologs, despite known interactions with the substrates. The *E. coli* WT sequence is indicated by green squares, while the average position fitness, fitness of a residue deletion, mean EVmutation evolutionary statistical energy [20], site conservation, relative solvent accessibility, and secondary structure information is shown above.**B.** The average fitness at each position, with blue and red representing low and high fitness respectively, overlaid on the *E. coli* PPAT (PDB: 1QJC, 1GN8 [21]) structure complexed with 4'-phosphopantetheine and ATP. We observe loss-of-function for mutations occurring at the active site, while other residues involved with allosteric regulation by coenzyme A or dimer interfaces show large promiscuity, highlighting different strategies employed among homologs. **C.** In addition to complementing homologs, we can also analyze mutants of the 129 low-fitness ($<$ -2.5) homologs, finding 385 gain-of-function (GoF) mutants across 55 homologs. We project this data onto the *E. coli* PPAT sequence and plot the number of GoF mutants at each position shaded by the number of different homologs represented. We find a total of 8 statistically significant positions (residues: 34, 35, 64, 68, 69, 103, 134, 135) corresponding to four regions in the PPAT structure.

did not complement. A total of 385 gain-of-function (GoF) mutants out of 4,658 were found for 55 homologs out of 129 low-fitness homologs (fitness $<$ -2.5). By aligning these mutations to the *E. coli* sequence, the eight statistically significant residues (34, 35, 64, 68, 69, 103, 134, 135) shown in Fig. 3.4C localize to four small regions in the protein structure (Fig. 3.26, Supplementary Text). We retrieved six GoF mutants of six different homologs from the library, each with fitness determined from only a single assembly barcode, and individually tested their growth rates. Five of the six mutants showed strong growth and one failed to complement (Fig. 3.21B). We also tested two of the corresponding low-fitness homologs, finding increases in the growth rate of 10% and 42% for their GoF mutants (Table 3.2).

Broad mutational scanning using DropSynth is a useful tool to explore protein functional landscapes. By analyzing many highly divergent homologs, individual steric clashes, which might be important to a particular sequence, become averaged across the homologs. More broadly, DropSynth allows for building large designed libraries of gene-length sequences, with no specialized equipment, and estimated total costs below $2 per gene (Table 3.3 & 3.4). We also show that DropSynth can be combined with dial-out PCR [15], which could be expanded for gene synthesis applications where perfect sequences are paramount. The scale, quality, and cost of DropSynth libraries can likely be improved further with investment in algorithm design, better polymerases, and larger barcoded bead libraries.

3.3 Materials and Methods

Design of PPAT library

PPAT homologs were found by running a PSI-BLAST search with 1 iteration querying the NCBI RefSeq non-redundant protein database using *E. coli* PPAT (NP_418091.1). The resulting set of 11,062 homologs was further pruned to 10,277 by keeping only those with lengths ranging from 100 to 200 amino acids. T-Coffee (v11) [24] was used to align the sequences and RAxML (v8.2.10) [25] to infer a maximum-likelihood phylogenetic tree (Fig. 3.6). A custom Python script trimmed the tree by determining the distance from the root at which the number of nodes equaled the desired amount of homologs, and subsequently choosing a random descendant leaf for each node at that

distance. This reduced the tree around 1,300 homolog proteins. Each leaf on the pruned tree was then compared to its nearest neighbours to ensure neighbouring sequences differed by at least five amino acids. We then added in homologs from several model organisms and 34 pathogenic organisms. The final library was dispersed among three libraries of 384 homologs, with every other leaf on the tree distributed into a different library, for a total of 1,152 homologs. The final library contained members sourced from 3 Archeal, 9 Eukaryotic, and 1140 Bacterial organisms (of which the top four most represented Bacterial phyla were 414 Firmicutes, 337 Proteobacteria, 64 Actinobacteria and 38 Spirochaetes).

Design of DHFR library

DHFR homologs were found using the DHFR family (IPR012259) in InterPro database [26]. A total of 5,760 homologs were selected, with 4,992 having lengths less than 530 bp (which can be assembled using 4X 230-mer oligos) and the rest greater (requiring 5X 230-mer oligos). Each homolog was encoded with two codon optimizations, creating a total of 30 libraries with 384 variants in each. We modified the oligo design scripts to forbid the presence of any homopolymer repeats greater than 8 nt. We also added random buffer sequence in between the KpnI restriction site and the reverse assembly primer to bring all of the assembled sequences to within 100 bp of each other to facilitate size selection. This extra buffer sequence is removed upon cloning of the assembled sequence into the barcoded vector. A single pool of 47,616 230-mer oligos was synthesized on a microarray by Agilent Technologies.

Microbead barcode design

We took 2,000 20-mer primers whose design was previously described [27] and removed those containing NdeI, XhoI, EcoRI, KpnI, NotI, SpeI, BtsI, or BspQI restriction sites. All possible 12-mer subset primers were generated and screened for self-dimers, GC content between 45% and 55%, and melting temperature between 40°C and 42°C. Barcodes were further filtered to ensure a minimum modified Levenshtein distance of 3 between any selected barcodes [28]. The first 384 12-mer barcodes, were used in subsequent oligo designs, with the complementary barcode sequences used to generate the beads.

Oligo design

Our oligo design protocol is adapted from Eroshenko et al [27] and summarized in Fig. 3.27A. Briefly, protein sequences were assigned a random codon weighted by the frequency in the *E. coli* genome, in order to generate a nucleotide sequence compatible with the restriction sites required (NdeI, KpnI, BtsI, or BspQI). A KpnI restriction site (*GGTACC*) was added on the C-terminal end of the coding sequence, which encodes for a glycine and threonine, before the stop codon. The NdeI restriction site (*CATATG*) on the N-terminal defined the start ATG codon of the ORF. Immediately flanking these restriction sites, 20-mer assembly primer sequences were added, which are used in the emulsion PCA. These sequences were then split into five shorter overlapping fragments [27], with overlaps optimized to be around 20 bp with a melting temperature between 58°C and 62°C. Sequences which failed to split with these parameters had a new weighted random codon assignment generated, until a codon sequence was found which could be split successfully. BtsI sites were subsequently added on either side of the split sequences, which would release the sequences required for assembly from the bead inside the emulsion droplets, allowing the PCA to proceed. A padding sequence consisting of ATGC repeats was added on the 5' end ahead of the first BtsI site, with the repeat length such that the final sequence length was 142 nt. Subsequently, an 8-nt Nt.BspQI site, the corresponding 12-mer microbead barcode (described above), and another Nt.BspQI site was prepended to the 5' end of the sequence, with the restriction sites oriented to nick the top strand on the 5' side of the barcode and the bottom strand on the 3' side of the barcode. These Nt.BspQI sites facilitate the processing of the barcode region into a single-stranded top-strand overhang. The barcode was common to all five fragments for each gene, such that all fragments required for each gene assembly would be pulled down and localized onto the same beads. Finally a pair of 15-mer amplification primer sequences were added, with each pool of 384 genes (1,920 oligos) having a unique primer pair orthogonal to the other pools. BLAT [29] was used to screen these primers against the oligo sequences, removing those with homologies over 10 bp. After each of these design steps, we screened for the addition of illegal restriction sites, and modified the sequence if any were found. For PPAT three libraries of 384 genes as well as a small test library of 24 genes were ordered as a single pool of 5,880 oligos (200-mer), while for DHFR thirty libraries of 384 genes were ordered as a single pool of 47,616 oligos (230-mer) and synthesized on a microarray by Agilent Technologies. For the PPAT libraries the final assembly

length was 52 bp longer than the gene length due to the addition of restriction and primer sites. The scripts required to generate DropSynth oligos are available at https://github.com/kosurilab.

DropSynth barcoded beads protocol

The general strategy for creating the DropSynth barcoded beads is shown in Fig. 3.27B. Three oligos are used for each DropSynth barcoded microbead, with two of the oligos common to all beads. The anchor oligo attaches to the streptavidin bead surface through a double biotin modification on the 5' end and has sequences necessary to hybridize with the ligation oligo and part of the barcode oligo. The ligation oligo has a biotin modification on the 3' end and phosphate group on the 5' which allows it to ligate to the microbead barcode oligo (Table 3.6). A different microbead barcode oligo is synthesized for each barcode with a common sequence on the 3' end which can hybridize to the anchor oligo and the reverse-complement of the microbead barcode on the 5' end which can pull down the gene fragments. This approach means only two synthesized oligos (anchor and ligation oligos) contain expensive modifications. Briefly the anchor oligo, ligation oligo, and each barcoded oligo are hybridized, ligated, and phosphorylated with T4 PNK. These are bound to streptavidin coated M270 Dynabeads, washed, and pooled together to form a uniform mixture of all 384 barcoded beads. This protocol can be scaled as necessary given the amount of multiplexing required. The current assembly protocol utilizes 18 uL of the final pooled bead mixture (∼3.25E5 beads/uL) for the capture of processed oligos, with the bead barcoding protocol provided producing enough pooled beads to carry out around 210 assemblies in 384-plex.

DropSynth protocol

DropSynth assembles gene-length fragments through the hybridization of oligos to barcoded microbeads and their resulting amplification. Briefly, individual oligo libraries are PCR-amplified using KAPA HiFi and 15-mer amplification primers. Oligo subpools are then bulk-amplified using the reverse amplification primer and a biotinylated forward amplification primer. After amplification, oligos are nicked using the nicking endonuclease Nt.BspQI, exposing a 12-nt ssDNA "barcode" overhang (Fig. 3.28, Table 3.5). The short biotinylated fragment that is cleaved following nicking is then removed by binding it to streptavidin M270 Dynabeads in a hot water bath. After a column

cleanup, each oligo subpool is mixed with the designed DropSynth barcoded beads and Taq ligase, and annealed overnight from 50°C to 10°C. In this process, all oligos required for each gene assembly are captured when each microbead barcode overhang anneals to a corresponding complementary microbead barcode on the bead. Captured beads are then mixed with KAPA2G Robust Mastermix, 20-mer forward and reverse assembly primers, BSA, BtsI, and BioRad Droplet Generation Oil. The mixture is immediately vortexed for 3 minutes, allowing for compartmentalization of captured beads in <5 um droplets (Fig. 3.29), which are subsequently heated allowing temperature-sensitive BtsI to release the sequences required for assembly from the bead. Droplets from each subpool are then loaded into PCR tubes and thermocycled, allowing PCA to proceed. The PCA products are then recovered by breaking the emulsion with chloroform, purified and re-amplified, providing sufficient assembled DNA for downstream applications.

Optimization of DropSynth

Significant optimization of the oligo processing and bead capture was required to achieve sufficiently high specificity to allow large multiplexing. Initial attempts to capture fully single-stranded oligos, generated using USER / λ exo / DpnII treatment [10], followed by primer extension of the missing complementary strand, performed poorly for three-oligo assemblies and failed altogether with four-oligo assemblies for all four polymerases tested (Kapa Robust, Kapa HiFi, Pfu Turbo, and Phusion). As an alternative approach, we nicked opposite strands on either side of the BC region with type IIS enzymes, before melting the microbead barcode strands apart and removing the unwanted biotinylated strand, leaving a single-stranded overhang along with the rest of the oligo, as shown in Fig. 3.1A. This eliminated the need for primer extension, and resulted in a 10-fold specificity improvement in tests on 96-plex assemblies of three to six oligos.

We also optimized the type of bead chemistry, testing both covalent carboxyl coupling and streptavidin coupling. Briefly, anchor oligos were covalently attached as follows. 100ul Dynabeads M-270 Carboxylic Acid were washed twice with 25 mM MES (pH 5). Next, 60μ g anchor oligo in 25 mM MES (pH 5) was added to the washed Dynabeads and incubated at room temperature for 30 minutes. EDC was dissolved in cold 100 mM MES (pH 5) to a concentration of 100 mg/ml, after which 30μ l EDC solution (3 mg) was added to the Dynabead/anchor oligo suspension. Next, 10μ l

of 25 mM MES (pH 5) was added and the solution was incubated overnight at 4°C with slow tilt rotation. Finally, the coated Dynabeads were washed 4 times using PBS (0.1% Tween-20). Despite successful assembly from carboxyl-coupled beads, we observed significantly higher enrichment factors in streptavidin-coupled beads. Thus we proceeded using streptavidin-coupled beads in all DropSynth experiments.

We further optimized the amount of beads, ligation reaction, the ligase used in the capture step, nicking reaction, presence/absence of size selection after assembly, and different techniques to purify the emulsion assembly products before re-amplification to achieve an assembly enrichment factor of 10^8, relative to the probability of a correct assembly by random chance, for a 288-plex five-oligo assembly (Fig. 3.1B).

PPAT rescue plasmid and coaD knockout

As PPAT (*coaD*) is an essential gene, we re-engineered plasmid pTKRED [30] and to constitutively express bicistronic wild-type (WT) *coaD* gene followed by sfGFP (Fig. 3.16A). The WT *coaD* gene from *E. coli* MG1655 was amplified with a strong constitutive promoter (TGACGGCTAGCTCAGTC-CTAGGTACAGTGCTAGC) and RBS (TACGAGTGAAAGAGGAGAAATACTAG) on the 5' end, and BamHI site on the 3' end. This was ligated to a fragment containing a 5' BamHI site, RiboJ self-splicing element [31], sfGFP [32], and a transcriptional terminator to create coaD_sfGFP. pTKRED was digested with BsaI and the larger fragment (8,391 bp) containing the $\lambda-$red genes was gel extracted. The coaD_sfGFP DNA fragment was then ligated into the larger pTKRED BsaI fragment to create pTKcoaD. This ligation was transformed into NEB 5-alpha electrocompetent *E. coli* and colonies were sequence verified. The pTKcoaD plasmid expresses PPAT and GFP constitutively while the $\lambda-$red recombinase genes are under IPTG induction. The temperature sensitive origin of replication can be used to heat cure the plasmid at 42°C, which can be confirmed through the loss of GFP fluorescence (Fig. 3.16E).

Knockout of the *coaD* gene in *E. coli* was carried out using standard techniques [30, 33]. Briefly pTKcoaD was transformed into both *E. coli* DH10B electrocompetent cells (ThermoFisher Scientific). Individual colonies were chosen and made electrocompetent. These were transformed with a recombination template containing a Kanamycin cassette flanked by homology arms to

the regions immediately adjacent to the *coaD* gene. This template was made by first amplifying the Kanamycin cassette from pZS2-123 [34] using primers coaD_KO_KAN_FWD_1 and coaD_KO_KAN_REV_1 (Table 3.7). The resulting amplicon was purified and further amplified using the primers coaD_KO_KAN_FWD_2 and coaD_KO_KAN_REV_2 (Table 3.7). The knock-in targeted only the PPAT coding region so as to not interfere with the essential waaA gene immediately upstream of *coaD*. Knock out strains were verified by Sanger sequencing and colony PCR (Fig. 3.16, C and D). We further verified that heat curing of the rescue plasmid suppressed cell growth and characterized the escape frequency.

pEVBC expression plasmid

The barcoded plasmid used to express PPAT homologs is a derivative of high-copy pUC19 with a pLac-UV5 promoter, NdeI and KpnI restriction sites for cloning, an in-frame stop codon, and a 20-mer random assembly barcode. This was made by first double-digesting pUC19 with AatII + BspQI and gel extracting the larger fragment. A gBlock DNA fragment was synthesized containing the promoter, several restrictions sites, and an in-frame chloramphenicol acetyltransferase before the stop codon. We initially tried using this in-frame chloramphenicol resistance as a way to screen the library against frame-shifted products, but we found this highly biased the resultant libraries (*data not shown*) and thus we did not use this in-frame selection for the results presented here. This was ligated into the pUC19 AatII-BspQI backbone fragment to create plasmid pEV_CMR. The plasmid pEV_CMR was double digested with NcoI + KpnI and the long 2,209 bp fragment was gel extracted. Round-the-horn PCR was carried out using 1 ng of the pEV_CMR digest as template, a forward primer pEVBC_FWD with a 5' biotin and a NdeI site, and a reverse primer pEVBC_REV1 with a 5' biotin (Table 3.7), a 20 N-mer random assembly barcode, and a KpnI site, for 5 cycles. This PCR product was further amplified with outer primers pEVBC_FWD and pEVBC_amp_FWD for 15 cycles (Table 3.7). This amplicon was column purified, digested with NdeI + KpnI, treated with rSAP, cleaned up with Streptavidin coated Dynabeads to remove the small fragments, and column purified again to create the vector pEVBC (Fig. 3.16B).

Barcoded PPAT library in pEVBC

Assembled PPAT homolog genes for each library were digested with NdeI + KpnI and column purified. A ligation was then carried out for each PPAT library using 150 ng of NdeI + KpnI digested pEVBC vector and 100 ng of digested PPAT homolog genes using 3,000U of T7 ligase in a total volume of 30 uL. This reaction was column purified and concentrated to a volume of 16 uL. NEB 5-alpha electrocompetent *E. coli* cells were then transformed using 3-4 uL of the purified ligation, resulting in over 10 million cfus per transformation. Overnight cultures grown in LB with Carbenicillin were miniprepped, quantified, and an equimolar pool from all three PPAT homolog libraries was created, henceforth referred to as sample S0.

Barcoded DHFR library in pEVBC

Analogous to PPAT, assembled DHFR homolog genes for each library were digested with NdeI + KpnI and column purified. A ligation was then carried out for each library using 150 ng of NdeI + KpnI digested pEVBC vector and 100 ng of digested DHFR homolog genes using 3,000U of T7 ligase in a total volume of 30 uL. This reaction was column purified and concentrated to a volume of 8 uL. In order to overcome known DHFR overexpression issues in *E. coli*, we directly PCR-amplified ligation products using primers mi3_FWD and mi3_REV_N7## (Table 3.7) to add p5 sequencing adapters and library indexes, rather than transforming and miniprepping.

Assembly barcode mapping

The assembly barcoded PPAT libraries were sequenced on two Illumina Miseq paired end 600-cycle runs, and DHFR libraries were sequenced on three Illumina Miseq paired end 600-cycle runs. Each library was PCR amplified using primers mi3_FWD and mi3_REV_N7## (Table 3.7) to add p5 sequencing adapters and library indexes. The resulting amplicons were size-selected using gel-extraction and quantified using an Agilent 2200 Tapestation. Samples were then pooled and sequenced on a Miseq using custom primers mi3_R1, mi3_R2, and mi3_index (Table 3.7). This resulted in 27,822,356 total reads (for PPAT) after merging the runs together. Barcode read counts for the S0 (unselected) library were generated by extracting the 20 bp sequence corresponding to the barcode region from the Read 2 sequences and using Starcode [35] to collapse barcodes within

a Levenshtein distance of 1 (Fig. 3.10). Sequencing data are available from the sequencing read archive (SRA) under BioProject PRJNA421181.

Briefly, the subsequent data processing was carried out as follows. All Fastq files had adapters trimmed in bbduk followed by paired-end read merging using bbmerge (from the BBTools package version 36.14). All reads were then concatenated and piped into a custom python script which generated a consensus nucleotide sequence for each barcode. The script works as follows. First, we split reads into the 20 nt assembly barcode and the corresponding variant, and generate dictionary that maps every assembly barcode to a list of variants associated to it. To eliminate assembly barcodes that are associated with two different variants, we calculate the pairwise Levenshtein distance of every variant associated with a given assembly barcode. If a certain percentage of these assembly barcodes (5%) are greater than a distance cutoff (10) then we consider the assembly barcode contaminated and drop it from further analysis. Finally, we generate a consensus sequence by taking the majority basecall at every position. Mapped consensus sequences were then translated until the first stop codon and sequences perfectly matched to any designed homologs were annotated.

Analysis of the number of reads per assembly barcode as a function of dilution revealed a small number of assembly barcodes with very high number of reads, as many as 300,000 by the fourth dilution, attributed to the emergence of adaptive mutations conferring a growth advantage at 42°C, which occur stochastically. We also deduce from the lack of GFP positive colonies in the plates at various steps in the dilution that these adaptive mutations did not occur in cells still harboring the rescue plasmid. A total of 18 barcodes from serial dilution replicate A and 16 barcodes from replicate B were removed from further analysis.

Mutant homolog sequences were annotated by first aligning the consensus nucleotide sequence for each barcode against the 1,152 designed PPAT homologs using bbmap. The resulting SAM file was parsed to extract the closest alignment match. A pairwise alignment of the amino acid sequence was carried out for each mapped barcode sequence (until the first stop codon) against its best PPAT homolog alignment match. Mutants within a distance of 5 amino acids from the designed sequence had their individual a.a. mutations annotated for further analysis downstream.

We estimated the number of chimeric assemblies computationally. First, we used a custom python script to divide our merged reads into 5 equally sized chunks. We then used BBMap (v

36.xx) to perform a pseudo-local alignment to a reference fasta containing all of our designed constructs. We refer to these alignments as pseudo-local as BBMap first searches for an optimal global alignment, and clips the reads if they return a higher score. We then tallied the number of chunks successfully aligned, as well as the number of different unique references each chunk aligned to. We then categorized each construct as follows:

- Perfect - all 5 chunks align to the same reference
- Chimeric - all 5 chunks align, but to more than one reference
- Possibly Chimeric - any number of chunks (not necessarily 5) align to more than one reference
- Junk - less than 5 chunks successfully aligned

BMS analysis

Briefly, we aligned all complementing homologs using MAFFT and created a lookup table for each residue of each homolog. For perfect homolog sequences we scanned through all residues and placed the homologs fitness into a BMS data table with the corresponding residue and E. coli position based on the alignment. For the mutants up to distance 5 a.a. from the perfect, we took only the mutated residues and added the fitness of the mutant into the BMS data table with the mutated residue and the corresponding E. coli position based on the alignment. For each residue and position in the BMS data table, the BMS fitness was determined as the median value of all of the corresponding data point at that position.

Classifier

We implemented a simple classifier to predict how different variants would perform in our assay. First, we categorized each variant into two bins based on whether or not their measured fitness score was greater than 0. We then selected 6 features for our model - the amino acid mutation, secondary structure class as assigned by DSSP (loop, beta-sheet, or alpha-helix), relative solvent accessibility as assigned by DSSP, sequence conservation, evolutionary coupling as predicted by EVMutation, and the frequency of residue substitution from the sequence alignment used for EVMutation's prediction. We used the R package Caret to perform a simple logistic regression using these features. To assess

the performance of our classifier, we performed 10 repeats of 5-fold cross-validation on our dataset and measured the precision and recall of each model on its respective hold-out set. We then used the R package precrec to plot both the receiver operating characteristic (ROC) and precision recall curves [36].

Complementation assay

The complementation of synthesized homologs was carried out using a serial batch culture. After ligation into pEVBC, homologs from all three libraries were pooled together to create sample S0. Supercoiled S0 plasmid was then electroporated into electrocompetent *E. coli* DH10β Δ *coaD* pTKcoaD. The serial batch cultures, consisting of two biological replicates, were initiated by making 10 transformations using 1 ng of S0 plasmid into 40 uL of cells and recovered at 30°C in 1 mL SOC + 1 mM IPTG for 1 hour. For each replicate, 5 transformations were pooled together and used to seed a fresh culture with between 7 million and 17 million cfus. Cells were grown in 1 L LB media supplemented with Kanamycin + Carbenicillin + 0.05 mM IPTG and grown to saturation at 42°C (8-10 generations) between each bottleneck. Cells were propagated through 3 bottlenecks for a total of 4 samples for each replicate, with 1000x dilutions at each bottleneck. DNA was miniprepped from each sample and cells were plated to ensure proper curing of the rescue plasmid, by screening for GFP+ colonies.

The barcodes from each of the 8 complementation samples were amplified using primers mi4_FWD and mi4_REV_N7## (Table 3.7) to add sequencing adapters and library indexes. The resulting 294 bp amplicon was size-selected using gel-extraction, purified, pooled, and loaded onto a Hiseq 2000 single-end 50 cycle run using custom sequencing primers mi4_R1 and mi4_index (Table 3.7), resulting in 138 million total reads. The barcodes for each sample were clustered using Starcode [35] to collapse barcodes within a Levenshtein distance of 1 (Table 3.1).

Complementation data analysis

In order to reduce noise in calculating the fitness change we pruned the barcodes leaving only those with at least 10 reads in S0 or at some point in the serial dilution. This reduced the total number of unique barcodes from 7,038,274 to 627,302. We calculated fitness scores for each mapped sequence

with at least one barcode. First, the read counts at each dilution were normalized based on the total sequencing depth of the sample relative to S0. The log2 fold change between each dilution and sample S0 was then determined for each barcode using

$f_{x0} = \log_2(r_x + 1) - \log_2(r_0 + 1)$,

where r_x is the number of normalized reads in the corresponding dilution. We then took the median value (to minimise effects of outliers) of the log2 fold change over all of the dilutions to determine the fitness for that barcode

$f_{\mathrm{BC}} = median(f_{10}, f_{20}, f_{30}, f_{40})$.

The median fitness for each barcode representing a sequence was determined for each replicate (A and B) individually

$f_{\mathrm{seqA}} = median(f_{BC1A}, f_{BC2A}, f_{BC3A}, f_{BC4A}, ...)$,

$f_{\mathrm{seqB}} = median(f_{BC1B}, f_{BC2B}, f_{BC3B}, f_{BC4B}, ...)$.

We then selected only those sequences represented in both replicates and took the median replicate fitness as the final fitness value

$f_{\mathrm{seq}} = median(f_{\mathrm{seqA}}, f_{\mathrm{seqB}})$.

Data analysis was carried out in R, with visualisations using ggplot2, ggtree [37], and UCSF Chimera [38]. Residue conservation was determined using Jensen-Shannon divergence [39], secondary structure and relative solvent accessibilities sourced from DSSP analysis [40, 41] of 1H1T [42]. The analysis scripts are available at https://github.com/kosurilab.

Assembly Retrieval by Dialout Amplification

The presence of a unique barcode on each assembly allows us to retrieve them from the library using PCR amplification [13, 15]. We attempted to amplify 48 unique homologs and 12 gain-of function mutants. As a positive control we also amplified the wild-type *E. coli* coaD gene from the pTKcoaD rescue plasmid. The designed primers flanked each construct, with reverse primers annealing to each gene-specific barcode. We observed correct size amplification products for 59 of 60, with 18 of these using lower complexity post complementation selection libraries as template, while the rest used the high-complexity sample S0. Individual amplicons were then gel-extracted, restriction digested with KpnI-HF and NdeI, ligated into empty pEVBC backbone, and transformed

into chemically competent NEB DH5alpha *E. coli* cells. Colonies were verified via colony PCR and Sanger sequencing, and validated colonies were re-inoculated overnight and miniprepped. We successfully sequence-verified 43 of the 59 constructs (37 homologs and 6 gain-of-function mutants), in addition to the WT coaD gene (Table 3.2).

Growth Rate Analysis of Dialed-out homologs and Gain-of-function Mutants

Following successful dialout PCR and re-cloning, we transformed 1 ng of each construct in pEVBC into 7 uL of electrocompetent *coaD* knockout cells. We analyzed the presence of growth by counting dilution Carb + Kan plates at both 30°C and 42°C (Table 3.2, Fig. 3.21A). Four constructs had no colonies on the 42°C plates, of which two were low-fitness homologs, one was a gain-of function mutant with only one barcode (false positive), and another construct KOS35328 had good fitness (1.88) in the pooled assay determined using 25 barcodes. The lack of colonies for KOS35328 requires further investigation, and may be a transformation error. Six constructs had low colony counts on both plates, of which five correspond to low-fitness homologs (Fig. 3.21A). We noticed a trend in which homologs with enhanced fitness in the pooled complementation assay gave rise to greater numbers of colonies on the 42°C dilution plates. Furthermore, we also noticed that homologs with enhanced fitness in the pooled complementation assay typically gave rise to 42°C colonies that appeared larger than their corresponding 30°C colonies (Fig. 3.21A). Of the constructs with at least 10 colonies on the 42°C plates, we picked 3 colonies per homolog and re-inoculated them in 1 mL LB + Carb + Kan and grew overnight at 42°C. 2 uL of saturated culture was then diluted in 98 uL of LB + Carb + Kan in wells of a 96-well plate and loaded into a Tecan M1000 Plate Reader for 12 hours at 42°C. OD600 values, taken at 30-minute intervals, were measured at 9 points within each well and averaged. Resultant growth curves were plotted for all colonies and averaged on the construct level. Maximum slopes of each growth curve were calculated and plotted against fitness scores determined from the complementation assay (Fig. 3.21B). A strong correlation (Spearman $r_s = 0.86$, p-value 5.9E-12) was observed comparing homolog growth rate to fitness, validating our assay and analysis pipeline. Examining the residual errors of the fit of growth rate to fitness we observe that constructs with fewer barcodes tend to have larger errors (Fig. 3.21C) which agrees with the reproducibility of the fitness value among replicates as a function of the number of barcodes

(Fig. 3.18B).

DropSynth bead barcoding protocol

Prepare 2X Binding and Wash buffer (2M NaCl, 1mM EDTA, 10mM Tris)

2X B&W 40mL:

- 4.675g NaCl salt
- 400 uL UltraPure 1M Tris-HCl, pH 7.5 (Invitrogen)
- 80uL UltraPure 0.5M EDTA, pH 8.0 (Invitrogen)
- UltraPure Distilled Water (Invitrogen) to 40 mL

This protocol can be done on a single 384 well plate or 4x 96 well plates, the latter protocol is provided. Reagents required:

- 384 uL 100 uM anchor oligo (Integrated DNA Technologies)
- 384 uL 100 uM ligation oligo (Integrated DNA Technologies)
- 1 uL 100 uM of each barcode oligo (Integrated DNA Technologies)
- 1,576 uL 10X T4 ligase buffer (New England Biolabs)
- 384 uL T4 PNK (10,000 U/mL) (New England Biolabs)
- 40 uL T4 ligase (concentrated 2,000,000 U/mL) (New England Biolabs)
- 1,920 uL stock Dynabeads M270 Streptavidin (Invitrogen)

For each of the four 96-well plates:

1. Mix 96 uL 100 uM anchor oligo and 96 uL 100 uM ligation oligo.
2. Prepare the 96 well plate. In each well add:
    - 2 uL of mixed anchor and ligation oligo

- 1 uL 100 uM barcoded oligo
- 4 uL 10X T4 Ligase buffer
- 33 uL UltraPure Distilled Water
- TOTAL: 40 uL

3. Anneal the mixed oligos on each plate using using the following conditions (30 min total):

    - 3 min at 70°C
    - Ramp down to 60°C for 1 min, 0.1°C/sec
    - Ramp down to 50°C for 1 min, 0.1°C/sec
    - Ramp down to 40°C for 1 min, 0.1°C/sec
    - Ramp down to 30°C for 1 min, 0.1°C/sec
    - Put plate on ice

4. Ligate the barcoded oligo to the ligation oligo:

    - Make a 1:10 T4 Ligase dilution:

    10 uL T4 Ligase (concentrated 2,000,000 U/mL)

    10 uL 10X T4 ligase buffer

    80 uL H_2O

    TOTAL: 100 uL

- Add 1 uL T4 Ligase (1:10 dilution) to each well
- Incubate plate at 16°C for 1 hr or longer, followed by 65°C for 20 min to heat inactivate the ligase

5. Phosphorylate the barcoded oligo:

    - Add 1 uL T4 PNK into each well

- Incubate the plate at 37°C for 40 min (or longer), followed by 65°C for 20 min to heat inactivate the PNK

5. Bind to beads:

   - Prepare 480 uL stock Dynabeads M270 Streptavidin, washed, and resuspended in 960 uL B&W buffer
   - Add 10 uL resuspended beads to each well. (∼3.25E6 beads/well and ∼18.5E6 molecules/bead)
   - Mix overnight with shaking (2000 RPM) at room temperature.

7. Pool beads:

   - Wash each well with 150 uL B&W buffer 5 times.
   - Resuspend in 10 uL B&W buffer
   - 1 uL of each well is mixed together, making a 96 uL mixed barcoded bead pool for each plate.
   - Mix 96 uL from each plate to make a full 384 uL mixed barcoded bead pool. Store these at 4°C when not in use.

DropSynth emulsion synthesis protocol

The following protocol was used to assemble the PPAT library. All PCR steps were performed on a Bio-Rad C1000 Touch Thermal Cycler (Bio-Rad Laboratories).

1. Prepare the OLS pool

   - Make 1/5, 1/10, and 1/20 dilutions of the OLS chip pool.
   - Prepare mixtures of forward and reverse subpool amplification primers for each subpool, with 10μ M final concentration of each primer.

2. Amplify subpools.

- For each subpool run a qPCR to determine the number of cycles required for amplification. Amplifications are stopped several cycles before plateauing to prevent over-amplification of the libraries.
- Amplify each subpool.
  - 1 uL template (test 1/5, 1/10, 1/20 OLS pool dilutions)
  - 1.25 uL subpool specific primer mix (10 uM FWD + 10 uM REV)
  - 22.75 uL UltraPure Distilled Water (Invitrogen)
  - 25 uL Kapa HiFi HotStart ReadyMix (2X) (KAPA Biosystems)
  - TOTAL: 50 uL
  - PCR protocol:
    1. 3 min 95°C initial denaturation
    2. 45 sec 98°C denaturation
    3. 15 sec 58°C annealing
    4. 15 sec 72°C extension
    5. Go to step 2, repeat based on the number of cycles determined by qPCR.
    6. 1 min 72°C final extension
- Column purify amplified oligos using a Zymo Clean & Concentrator -5 (Zymo Research).
- Run PCR products on gel. Look for higher MW products, indicative of overamplification. Excessive low MW products may indicate chip synthesis issues.
- Size select, using gel extraction, if necessary.
- Create 20 pg/uL dilutions of each amplified subpool. (∼91 million/uL)

3. Bulk amplify subpools.

- Run a second PCR using a *biotinylated* FWD amplification primer, with sufficient tubes to make 4.5 ug to 9 ug of PCR product.
  - 1 uL of 20 pg/uL subpool dilution
  - 1.5 uL subpool specific primer mix (10 uM biotinylated FWD + 10 uM REV)

- 22.5 uL UltraPure Distilled Water (Invitrogen)
- 25 uL Kapa HiFi HotStart ReadyMix (2X) (KAPA Biosystems)
- TOTAL: 50 uL
- PCR protocol:
  1. 3 min 95°C initial denaturation
  2. 20 sec 98°C denaturation
  3. 15 sec 58°C annealing
  4. 15 sec 72°C extension
  5. Go to step 2, 18X
  6. 1 min 72°C final extension

- Pool and column purify PCR reactions using a Zymo Clean & Concentrator -5 (Zymo Research).

4. Nicking.

   - Nick the bulk amplified subpools. Split the following accross multiple tubes depending on the amount of DNA to be processed. In each 1.5 mL tube add:
     - 4.5 uL Nt.BspQI (10U/uL) (New England Biolabs)
     - 2 to 2.5 ug of DNA (final concentration ∼16ng/uL)
     - 15 uL NEBuffer 3 (New England Biolabs)
     - UltraPure Distilled water (Invitrogen) to 150 uL
   - Leave at 50°C overnight with shaking >1500 RPM.

5. Capture and remove the short biotinylated fragment.

   - Wash 50 uL Dynabeads M-270 Streptavidin (Invitrogen) for each 1.5 mL tube in the nicking reaction, as per manufacturer's instructions and resuspend in 2X B&W buffer.
   - Add 50 uL of washed beads to the 150 uL nicking reaction in each tube.
   - Incubate at 55°C with 800 RPM shaking for at least 1 hour.

- Move all 1.5 mL tubes to a 55°C water bath.
- Place the tube so that solution is just below the surface of the water. Hold a strong magnet underwater against the side of the tube to magnetically separate Dynabeads. Pipette the supernatant, which contains the processed oligos and save them in a new container. Remove the tube with the Dynabeads from the magnet. Add 100 uL of UltraPure Distilled water (Invitrogen) to the tube and resuspend the beads. Incubate these at 55°C for another 30 min and then repeat the procedure to recover the supernatant again while leaving the Dynabeads behind.
- Repeat this procedure for all tubes as necessary.
- Pool processed oligos (supernatant) for each subpool and column cleanup using a Zymo Clean & Concentrator -5 (Zymo Research).

6. Capture processed oligos with barcoded beads.

- Take 18 uL of the pooled barcoded beads. These are in stored in B&W buffer (high ionic concentration) which may interfere with ligation reaction. Resuspend them in 18 uL 10 mM Tris-HCl buffered solution.
- Mix the processed DNA with the barcoded beads:
  - 40 uL processed DNA (∼1.3 ug,∼12 pmol)
  - 18 uL pooled barcoded beads (∼5 million beads, binding capacity 1.2 ug DNA)
  - 10 uL 10X Taq ligase buffer (New England Biolabs)
  - 4 uL Taq ligase (40 U/uL) (New England Biolabs)
  - 28 uL UltraPure Distilled water (Invitrogen)
  - TOTAL: 100uL
- Overnight cycling (>2 hr incubation at each of the following temperatures) (13 hr), while shaking using an Eppendorf ThermoMixer C (Eppendorf):
  - 3 hours @ 50°C
  - Ramp to 40°C for 3h, 0.1°C/sec

  - Ramp to 30°C for 3h, 0.1°C/sec
  - Ramp to 20°C for 2h, 0.1°C/sec
  - Ramp to 10°C for 2h, 0.1°C/sec
- Wash 3 times at 4°C using B&W buffer. This is important for removing unbound oligos in order to increase specificity.
- Wash twice at RT using B&W buffer
- Re-suspend in 100 uL Elution Buffer (Qiagen) (∼50k beads/uL)

7. Emulsion assembly (ePCA).

- Setup emulsion. All of this procedure should be done in cold room on ice. Add Bts α I only at very last step. Try to minimize the time between adding the Bts α I and vortexing the emulsion.
  - 10 uL of loaded beads (∼130 ng DNA)
  - 0.5 uL 100 uM FWD assembly primer
  - 0.5 uL 100 uM REV assembly primer
  - 50 uL Kapa2G Robust HotStart ReadyMix (2X) (KAPA Biosystems)
  - 1 uL BSA (New England Biolabs)
  - 31 uL UltraPure Distilled water (Invitrogen)
  - 7 ul Bts α I (New England Biolabs) (*add last*)
  - TOTAL: 100 uL
- Mix at low speed in vortexer to resuspend beads.
- Add 600uL Droplet Generation Oil for EvaGreen (Bio-Rad Laboratories) to a 1.5mL non-stick tube.
- Add 100uL aqueous phase to the bottom of the oil phase.
- Vortex at Max Speed in foam holder taped down for 3-4 minutes. If doing multiple emulsions, do this one at a time. We use a Vortex Genie 2 (Scientific Industries) at max speed.

- After vortexing all emulsions, place each emulsion into PCR tubes with 100 uL in each tube. Use a P1000 tip to avoid disturbing the emulsion. Most of the droplets will float to the top of the tube, try to get as much of this as possible and distribute this over multiple PCR tubes.
- PCR Cycling
  - 55°C for 90 min (allow Bts α I to cleave DNA from the beads)
  - 94°C for 2 min (initial denaturing)
  - 94°C for 15 sec (denaturing)
  - 57°C for 20 sec (annealing)
  - 72°C for 45 sec (extension)
  - Go to step 3 for additional 60 cycles
  - 72°C for 5min (final extension)
  - 4°C forever

8. Break the emulsion. Adapted from pg 69 of the Bio-Rad Droplet Digital PCR Applications Guide:

   - After ePCA, pipet out the entire volume of droplets from each PCR tube into a 1.5 mL tube. Combine up to 400 uL, in each tube. Note: phase-lock tubes can also be used here to improve recovery.
   - Carefully pipet and discard bottom oil phase after droplets float to the top. Press a P1000 down to its first stop, push through the droplets to the bottom of the tube, press down to the second stop to expel any droplets, then wait several seconds for the droplets to float back up to the droplet layer, and finally aspirate out the oil. You do not need to remove every last bit of oil - just remove most of it.
   - Add 50 uL of TE buffer for each 100 uL of PCR reaction combined in the 1.5mL tube.
   - In a fume hood, add 175 uL of chloroform for each PCR reaction in the tube. (If there are 4 PCR reactions in a tube than contents will be: <400uL PCR reactions, 200uL TE, 700 uL chloroform).

- Vortex at maximum speed for 1 min.
- In a centrifuge, spin down at 15,500 x g for 10 min.
- Remove upper aqueous phase by pipetting, avoiding the chloroform phase.
- Transfer this to a clean 1.5mL tube (this is the DNA).
- Proceed to column or SPRI bead cleanup (Beckman) for the recovered DNA.

9. Size selection.

- The amplicons will often be mixed with undesired lower-molecular weight assemblies. Removing these using size selection will increase final yield. Choose of of the following three approaches, ordered from highest yield to lowest yield:
  - Pippin Prep (Sage Science).
    1. Follow manufacturer's protocol (calibration, checking currents, loading, etc. . .)
    2. Make sure to allow for a range broad enough to include every member of the library, yet narrow enough to exclude some of the shorter non-specific products (+/- 100 bp is usually fine).
    3. Collect the eluted product and column cleanup using a Zymo Clean & Concentrator -5 (Zymo Research).
  - or Gel extraction.
    1. Run amplicons on a gel and extract the correct range and purify.
    2. Note: Typically there is not enough DNA after the ePCA to visualize on a gel, so this is often a blind extraction.
  - or No size selection.
    1. Make a dilution of ePCA and use this as template for the re-amplification.

10. Re-amplification.

- Amplify ePCA products using Kapa HiFi HotStart ReadyMix (2X) (KAPA Biosystems).
  - 0.2 - 2 uL template

- 1 uL 10 uM FWD assembly primer
- 1 uL 10 uM REV assembly primer
- 25 uL Kapa HiFi HotStart ReadyMix (2X) (KAPA Biosystems)
- UltraPure Distilled Water (Invitrogen) to 50 uL
- TOTAL: 50 uL
- PCR protocol:
  1. 3 min 95°C initial denaturation
  2. 15 sec 98°C denaturation
  3. 20 sec 58°C annealing
  4. 45 sec 72°C extension
  5. Go to step 2, determine cycles using qPCR.
  6. 3 min 72°C final extension

- Column purify re-amplified products using a Zymo Clean & Concentrator -5 (Zymo Research).
- Check size distribution on gel or tapestation.
- Quantify DNA and proceed to downstream applications.

Supplementary Text

PPAT complementation assay

There are several reasons homologs could have low fitness including environmental mismatches, improper folding, mismatched metabolic flux, interactions with other cytosolic components, or gene dosage toxicity effects resulting from improperly high expression. Of the homologs from extremophilic bacteria, only alkaliphiles showed slightly reduced fitness values which is not significant (p-value = 0.059 Wilcoxon) (Fig. 3.23B). Metabolic mismatch is unlikely since so many homologs were able to complement well and both CoA and dephospho-CoA act as inhibitors implementing negative feedback loops to control the metabolic flux through the pathway [17]. Control experiments revealed that high expression levels of wild-type *E. coli* PPAT result in growth defects, while similar levels of

expression for many other homologs had no impact (Table 3.2). This observation parallels similar findings for *E. coli* DHFR where wild-type overexpression was toxic while overexpression of homologs had no detrimental effects [19], an effect linked to evolved protein-protein interactions that confer benefits at physiological concentrations. PPAT interaction partners include several enzymes encoded by essential genes such as *leuS*, *murE,* and *rplD* [43, 44].

Broad Mutational Assay (BMS)

Although 87% of the 3,180 possible mutations are covered, the coverage is strongly correlated with position fitness (ρ=0.76; Pearson, p-value <2.2E-16) (Fig. 3.24C), implying that many mutations that are depleted in the pooled assay (and typically represented by a only a few assembly barcodes), never pass the 10-read threshold used to filter assembly barcodes, an issue that can be resolved by sequencing the initial library to a greater depth. Unlike traditional mutagenesis approaches, the presence of multi-bp deletions from the oligo synthesis process also allows us to evaluate the effect of removal of entire residues from the sequence (Del. in Fig. 3.4A).

Gain-of-Function Mutants

In *E. coli*, residue Glu-134 and proximal Leu-102 have hydrophobic interactions with the cysteamine moiety of CoA [42], suggesting that some GoF mutations play roles in tuning CoA inhibition, while Ala-103 participates in hydrophobic interactions contributing to dimer formation [17]. Residues 64, 68, 69 are surface-exposed in the hexameric PPAT complex and are possible candidates for interactions with other proteins. As many of these mutations had only a single assembly barcode, we estimated a false positive rate of 0.9% derived from the number of positive fitness mutants for negative controls (Fig. 3.19A).

Acknowledgements

This work was supported by the funds from the Human Frontier Science Program [LT000068/2016 to C.P.], Netherlands Organisation for Scientific Research Rubicon fellowship [to C.P.], National Science Foundation Graduate Research Fellowship under Grant No. 2016211460 [to A.M.S], a Ruth L. Kirschstein National Research Service Award [GM007185 to N.L.], National Institutes of Health

New Innovator Award [DP2GM114829 to S.K.], Searle Scholars Program [to S.K.], Department of Energy (DE-FC02-02ER63421 to S.K.), UCLA, and Linda and Fred Wudl. We thank Jeff Sampson and Paige Anderson at Agilent Technologies for oligo pools and critical advice. We thank George Church and Richard Terry for guidance during the early developments and Suhua Feng, the UCLA BSCRC Sequencing Core, and the Technology Center for Genomics & Bioinformatics for providing NGS services. S.K. and D.Z. are named inventors on a patent application on the DropSynth method (US14460496). The scripts required to generate DropSynth oligos are available at `https://github.com/kosurilab/DropSynth`. Sequencing data are available from the sequencing read archive (SRA) with the accession number SRP126669.

3.4 Supplementary Information

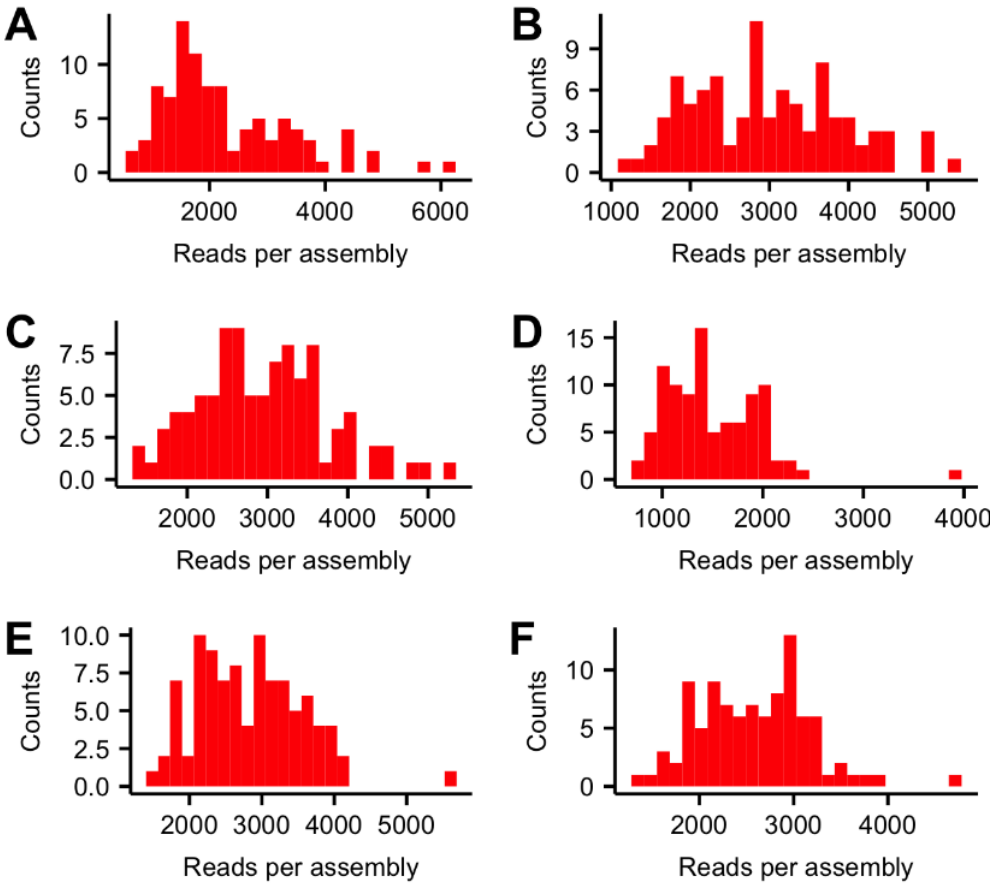

Figure 3.5: **The histogram of read distributions for six of the 96-plex 4-oligo assemblies shown in Fig 1B. A.** T7 ligase and 20 ug beads. **B.** T4 and 20 ug beads. **C.** Taq ligase and 20 ug beads. **D.** T7 ligase and 100 ug beads. **E.** T4 ligase and 100 ug beads. **F.** Taq ligase and 100 ug beads.

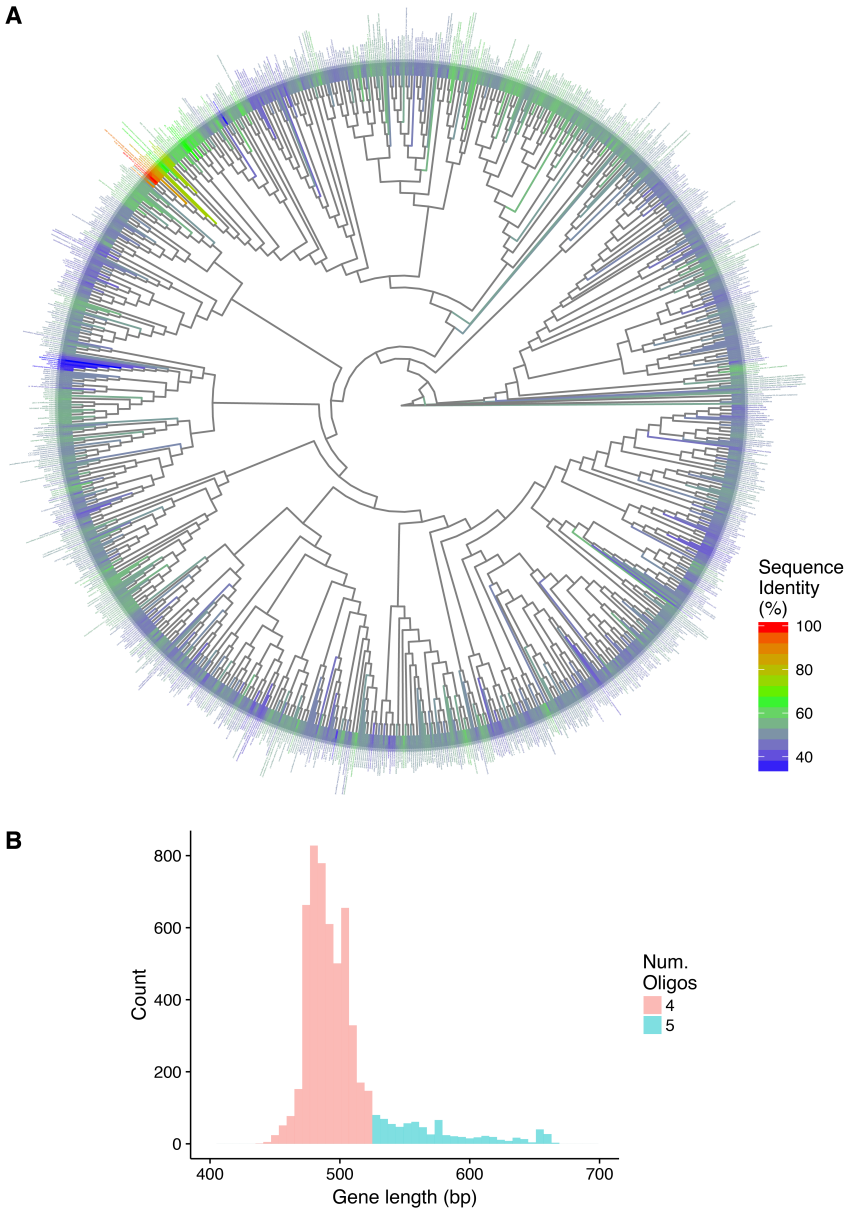

Figure 3.6: **A.** A maximum likelihood phylogenetic tree for all 1,152 PPAT homologs as well as *E. coli* MG1655. Color scale represents percent amino acid sequence identity relative to E. coli PPAT (NP_418091.1). **B.** The gene length distribution for the 5,775 DHFR homologs assembled using either four or five 230-mer oligos with median gene lengths of 489 bp and 564 bp respectively.

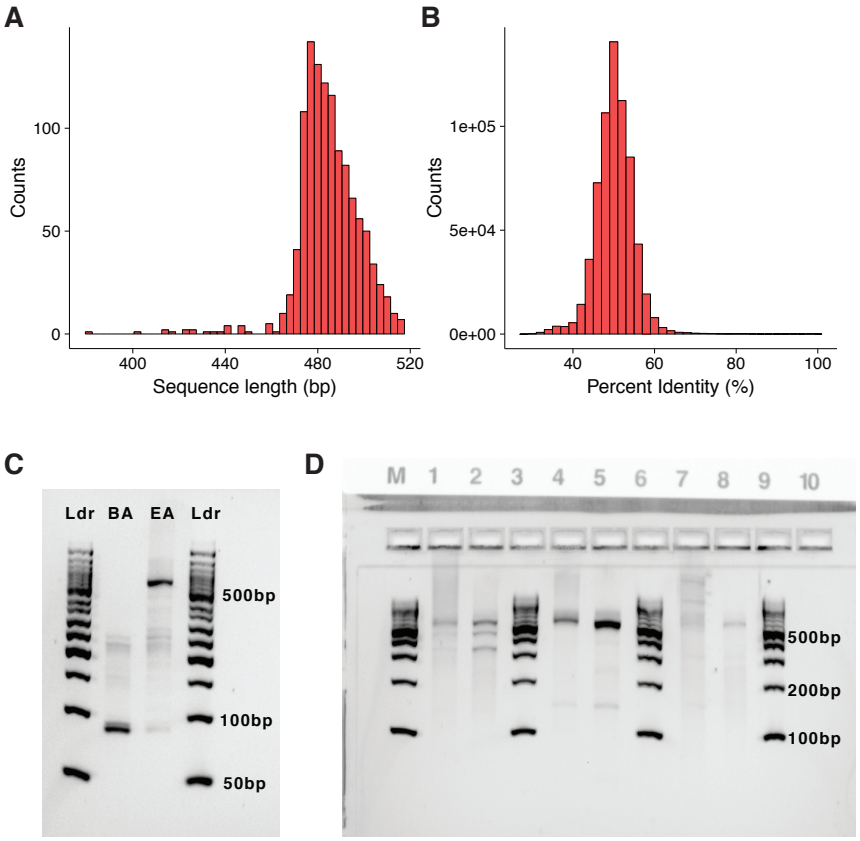

Figure 3.7: **A.** Histogram of protein sequence lengths for all 1,152 PPAT library members. Lengths do not include start or stop codon. The longest, shortest, and median lengths are 516, 381, and 483 bp respectively. **B.** Although they share the same function, PPAT homologs have evolutionarily divergent sequences. The 662,976 pairwise percentage identities between the 1,152 members of the PPAT library at the amino acid level have a distribution with a median of 50% ($\sigma = 5\%$). **C.** Without oligo isolation, amplification in bulk fails to produce the correct product [11]. A 4% agarose gel comparing the assembly products of a 24-member library of PPAT homologs (120 oligos) when the polymerase cycling assembly is done in bulk (BA) and in emulsion (EA). The expected product size upon correct assembly is between 520 bp to 550 bp. **D.** Each of the three 384-member PPAT libraries (1,920 oligos each) produced correct assembly products. A 4% agarose gel showing amplified assembly products, with the expected size for most amplicons around ∼530 bp. Lane 1 and 2: High- and low-template PCR products for Lib 1. Lane 4 and 5: High- and low-template PCR products for Lib 2. Lane 7 and 8: High- and low-template PCR products for Lib 3. High- and low-template concentrations refer to either 2 uL or 0.2 uL of the purified assembly products from an emulsion used in a 50 uL PCR reaction.

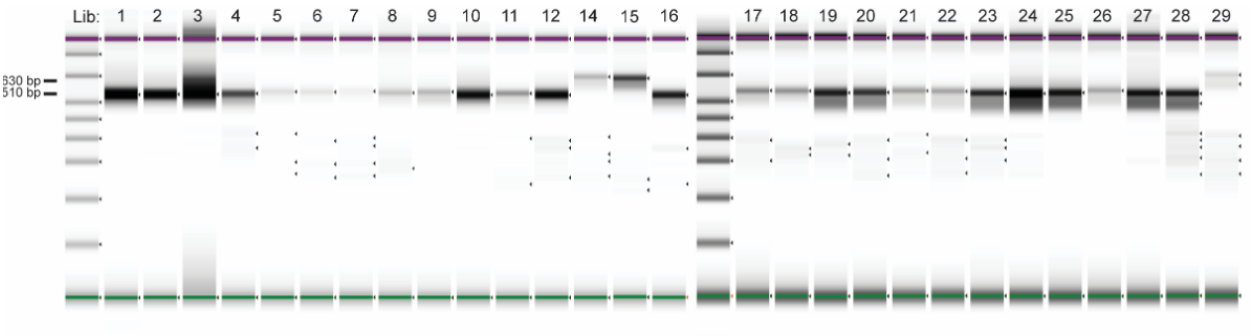

Figure 3.8: **Agilent TapeStation gel image of DropSynth assembly of 28 384-member libraries of DHFR.** A total of 3 libraries of length 610bp (14, 15, 29) are assembled using 5 oligos while the remaining libraries of length 510bp are assembled using 4 oligos. Another 2 libraries (13, 30) are not shown with one having low yield on the oligo processing steps and another failing to amplify at the oligo stage.

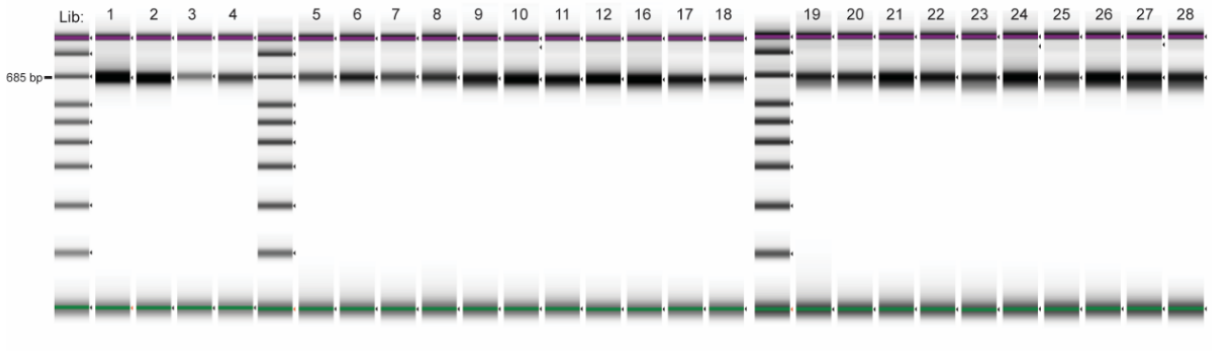

Figure 3.9: **Agilent TapeStation gel image of 25 4-oligo DHFR libraries after assembly, digestion, ligation into barcoded plasmid and library preparation for sequencing.** 5-oligo libraries (14, 15, 29) were not prepared for sequencing due to limitations on Illumina read length capabilities.

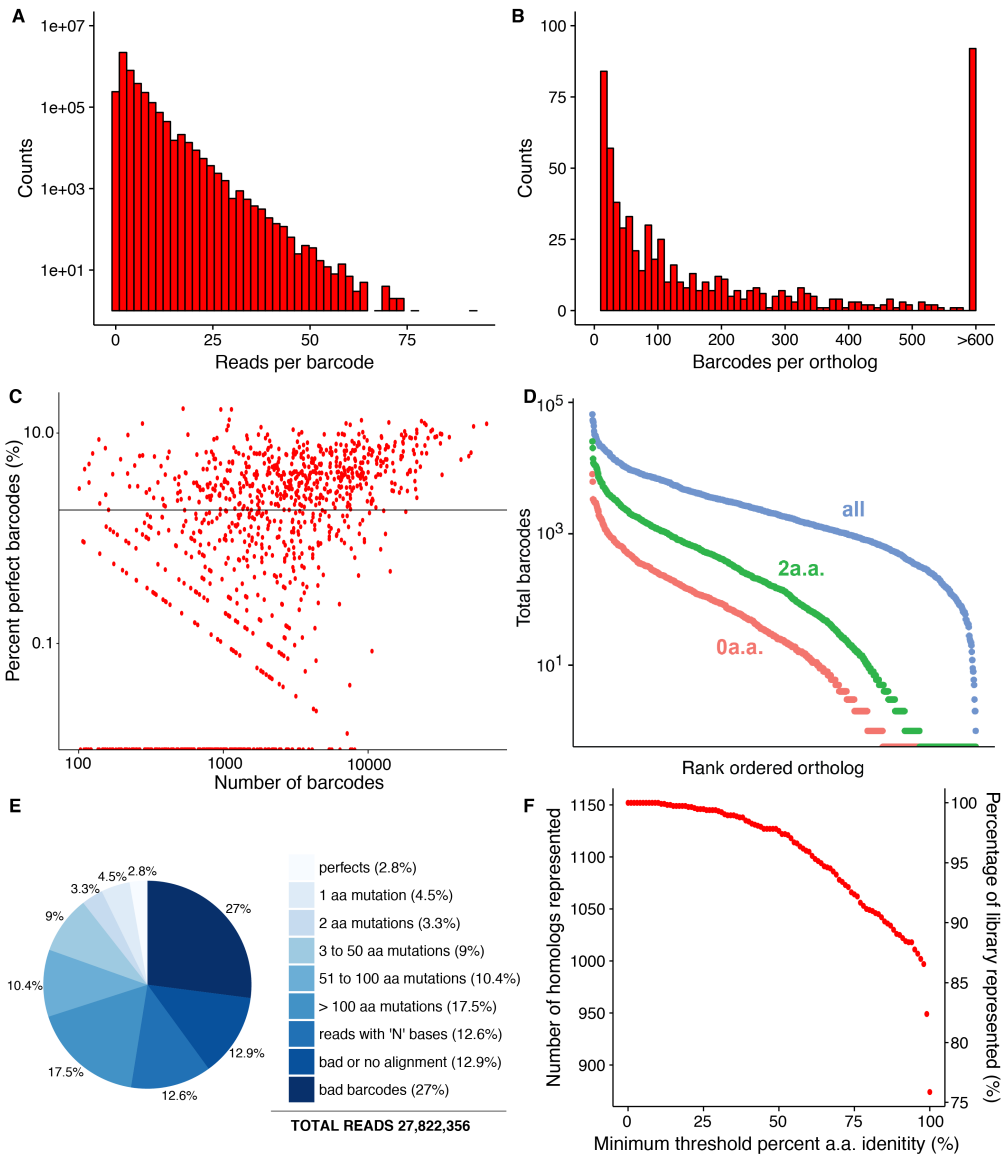

Figure 3.10: **Sequencing statistics from sample S0.** These data are a set of paired end 600-cycle Miseq runs which read through the entire assembled gene and its assembly barcode for all three 384-member libraries. **A.** The number of reads per assembly barcode, with a median value of 2. S0 contains 7,038,274 unique assembly barcodes across 20,263,445 reads. Of these, 209,868 assembly barcodes 2.98% (739,771 reads 3.65%) mapped to the designed protein sequences without any amino acid mutations, of which 199,208 assembly barcodes contained at least one synonymous mutation. A total of 2,982,539 (42%) of the mapped assembly barcodes correspond to sequences containing a premature stop codon in the reading frame, of which the large majority (2,404,348) were due to indel mutations causing a frameshift while the rest were due to nonsense mutations. **B.** The long tail distribution of assembly barcodes per homolog, for assembly barcodes mapped to a perfect sequence. Median value is 56 and a total of 872 out of 1152 homologs are represented with at least one assembly barcode. **C.** The percentage of perfect protein sequences for constructs with at least 100 assembly barcodes. The solid line is the median value of 1.9%. **D.** Individually rank-ordered plots showing the number of barcodes with perfect assemblies, barcodes with assemblies within distance of 2 a.a., and all barcodes with an aligned homolog. **E.** The distribution of sequencing reads for the PPAT libraries. **F.** The coverage of the PPAT homologs as a function of the minimum percent identity. Most of the library members have assemblies with high identity to the respective designed homologs.

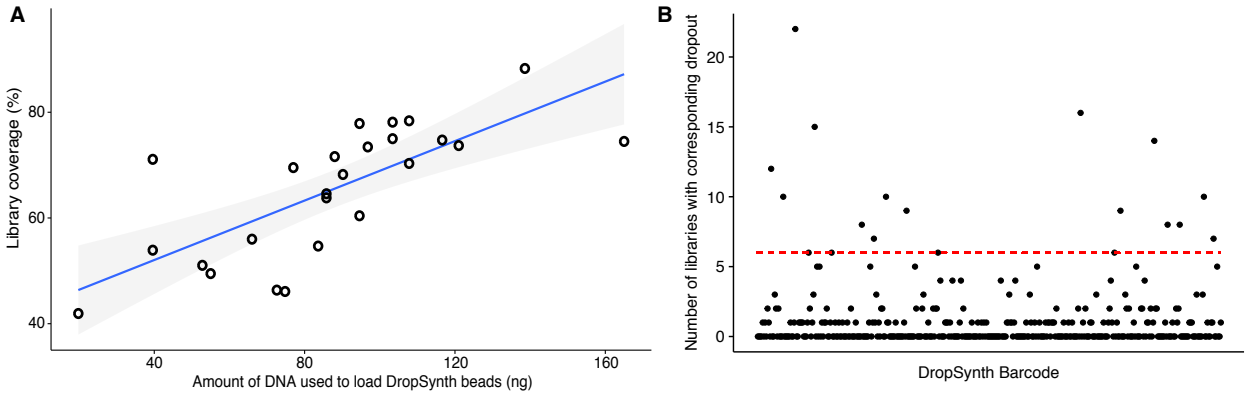

Figure 3.11: **A.** The library coverage shows strong correlation (ρ=0.73 (Pearson), p-value=3.4E-5) with the amount of DNA used to load the DropSynth beads prior to assembly. The coverage is defined as the number constructs with at least one perfect assembly. **B.** The number of constructs with the same barcode which dropout among different libraries. The red line is the level with an expectation value close to one for libraries of size 384 given a uniform dropout distributions. Values above this line are higher than would be expected by chance. About a dozen barcodes fall in this region.

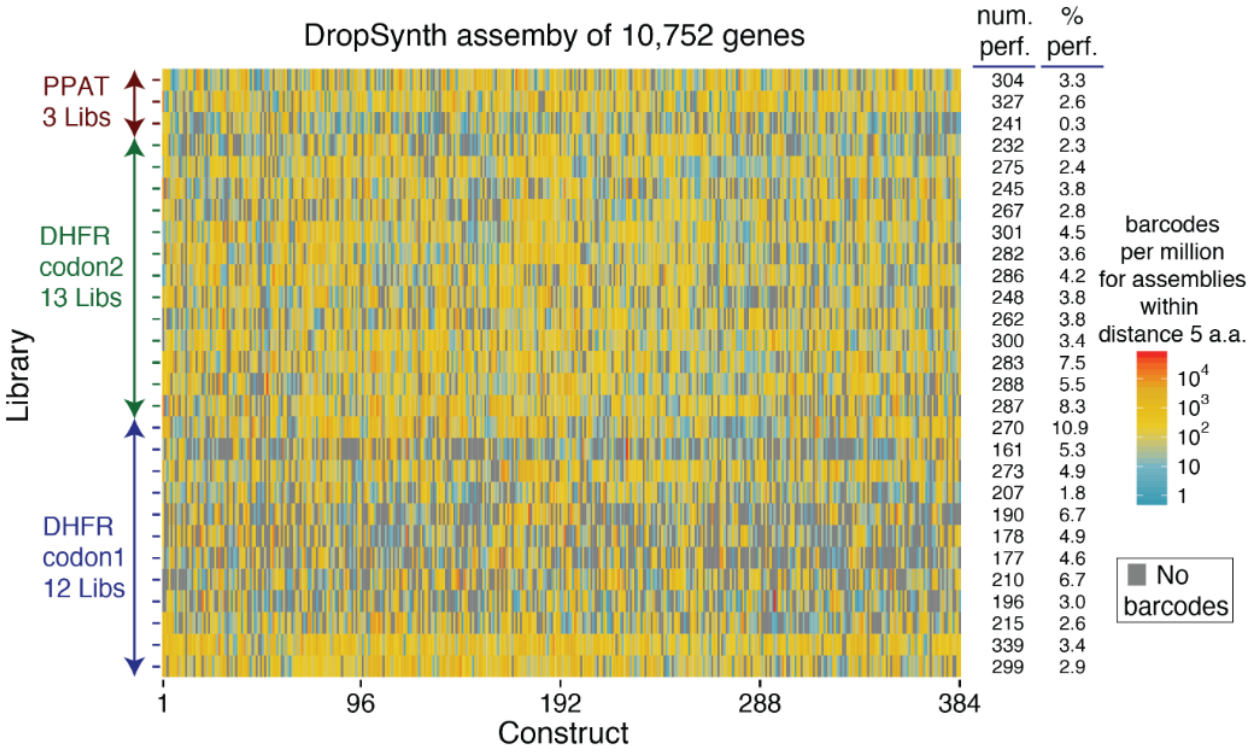

Figure 3.12: **DropSynth assembly of 10,752 genes.** We used DropSynth to assemble 28 libraries of 10,752 genes representing 1,152 homologs of PPAT and 4,992 homologs of DHFR. The number of barcodes per million representing assemblies within 5 a.a. of each gene is shown alongside the number of library members with at least one perfect assembly and the percent perfects determined using constructs with at least 100 barcodes.

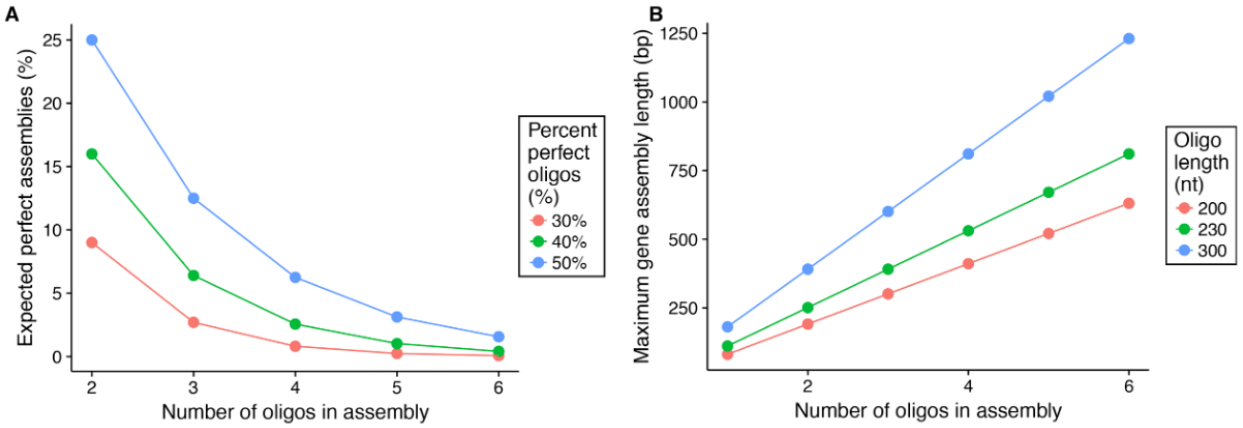

Figure 3.13: **A.** The expected percentage of perfect assemblies for a given number of oligos and the amount of perfect oligos. **B.** The maximum gene assembly length possible for a given number of oligos and an oligo size ranging from (200 to 300bp).

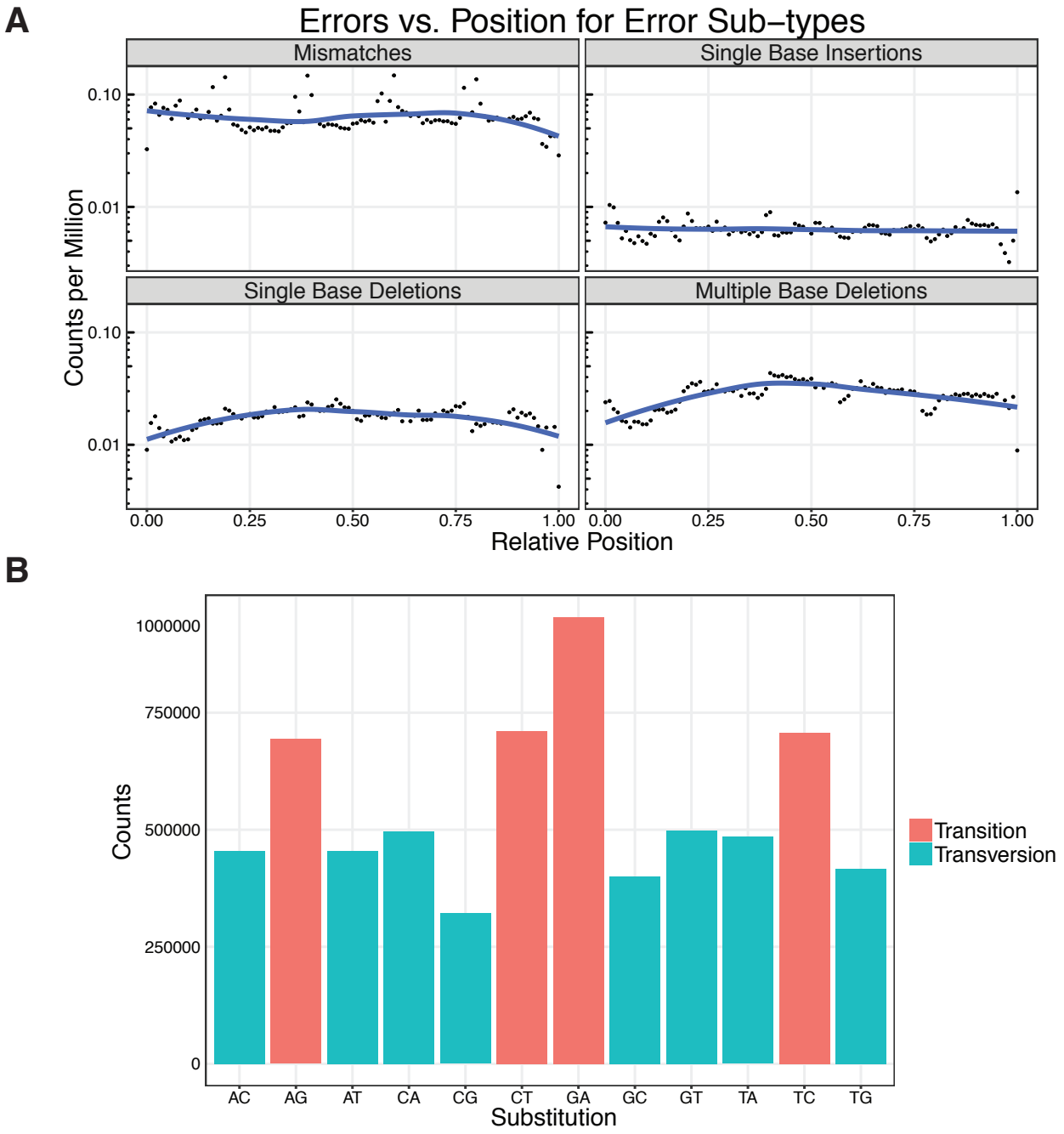

Figure 3.14: **Error analysis of DropSynth Assemblies.** Using the error analysis pipeline developed by Lubock et. al [16], we randomly sampled one million reads from Miseq paired-end 600-cycle assembly barcode mapping data, performed an exhaustive alignment of each read against every perfect assembly and returned the best scoring alignment. **A.** Mismatches are the most common form of error, followed by multiple base deletions, single base deletions, and single base insertions. In particular, mismatches appear to be localized to the overlap regions. **B.** Raw counts of mismatches. A higher number of transitions than transversions were measured - in agreement with previous experiments where Taq-mediated amplification errors. This suggests that the majority of mismatches were likely introduced by KAPA2G Robust polymerase during assembly (evolved Taq variant).

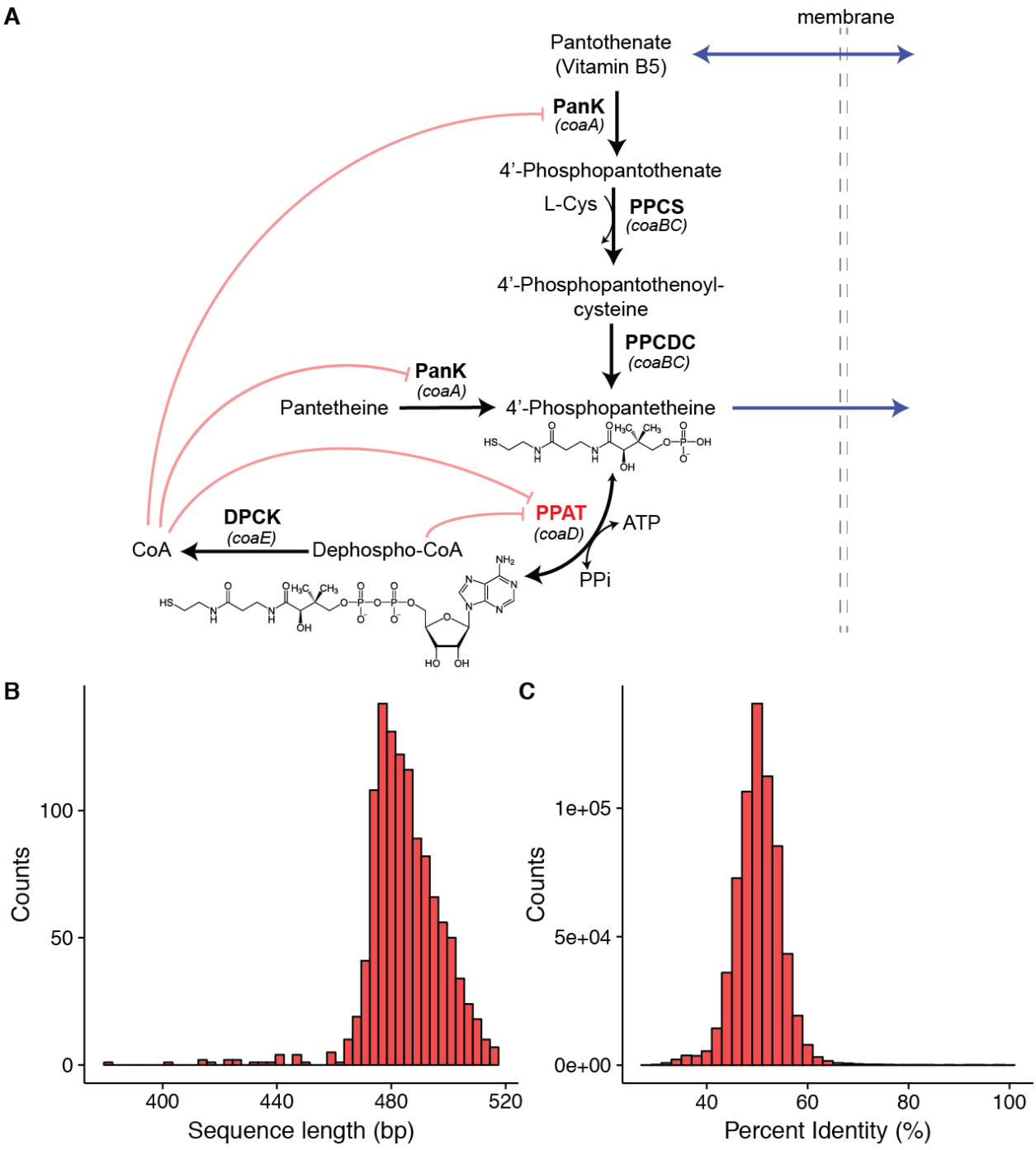

Figure 3.15: **Phosphopantetheine adenylyltransferase (PPAT) metabolic pathway.** PPAT shown in red, catalyzes the second to last step in the five step biosynthesis of coenzyme A. It produces dephospho-coenzyme A from 4'-phosphopantetheine by transferring a adenylyl group from ATP [17], as shown. Either Mn^{2+} or Mg^{2+} acts as a cofactor. *E. coli* PPAT is hexameric and encoded by the 477 bp gene *coaD*. Several gene knockout [45, 46] and genetic footprinting [47] studies have confirmed *coaD* to be essential for growth on rich media in *E. coli* K-12 strains MC1061, MG1655, and DH10β. Both coenzyme A and dephospho-coenzyme A act as inhibitors of the forward reaction. PPAT's low homology to its mammalian counterpart, which is encoded as one of the two domains on the bifunctional CoASy (CoA Synthase) enzyme, makes it a potential target for new antimicrobials [18]. At least a dozen different PPAT homologs have crystal structure data available.

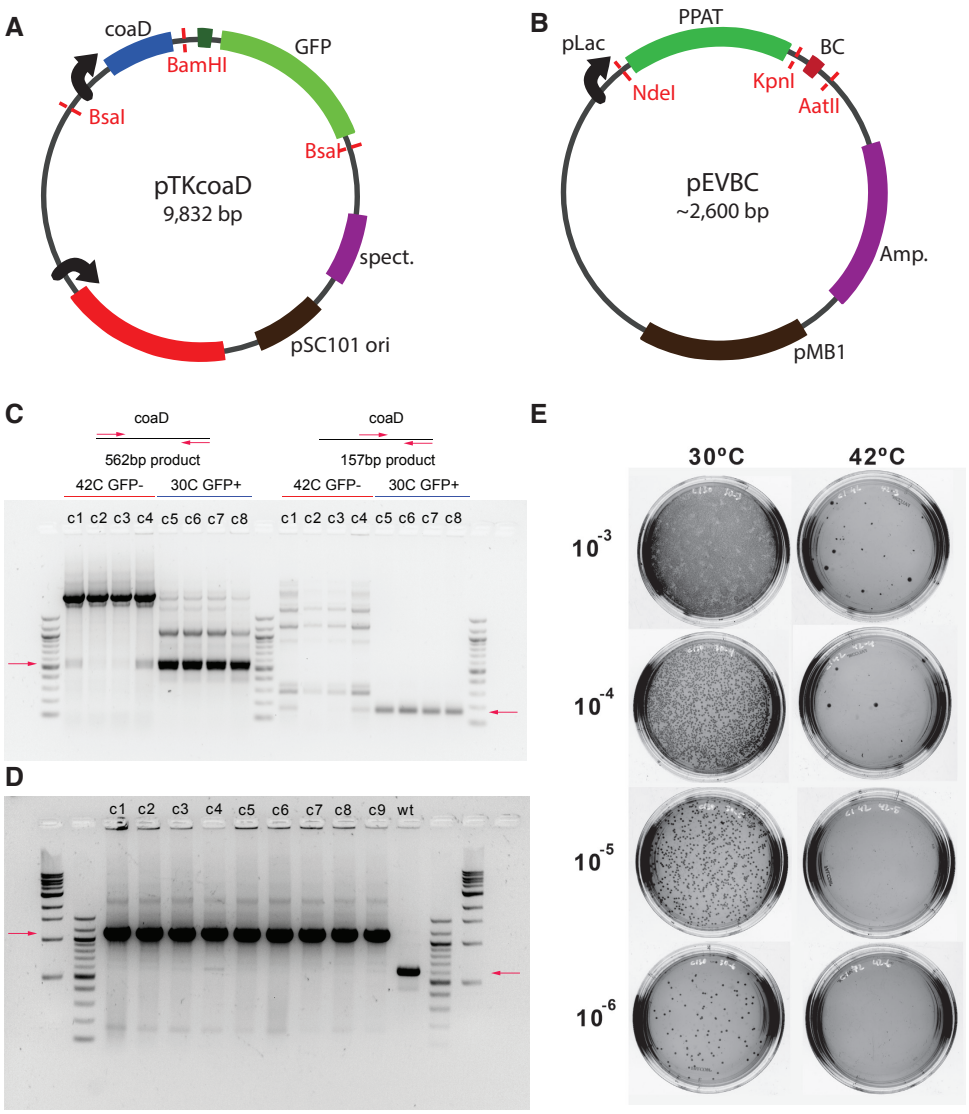

Figure 3.16: **A.** Rescue plasmid pTKcoaD allows λ−red recombination of the essential *coaD* gene. Wild-type *E. coli coaD* is expressed constitutively along with GFP, which allows for confirmation of plasmid loss upon heat curing. **B.** High-copy expression plasmid pEVBC allows for IPTG-inducible expression of an homolog PPAT gene cloned in between the NdeI and KpnI sites. A 20-mer random assembly barcode is present downstream. **C.** Verification of the *coaD* gene knockout using colony PCR with two sets of internal primers. Four 42°C heat-cured colonies (c1-c4) are shown as well as four colonies (c5-c8) grown at 30°C which still contain the rescue plasmid. Red arrows indicate expected amplicon size when *coaD* gene sequence is present. **D.** Colony PCR verification of the *coaD* genomic knockout using external genomic primers for 9 knockout colonies and one wildtype control. Wildtype (no knockout) amplicon length is 590 bp while the knockout (KAN cassette knockin) amplicon length is 1150 bp, as marked by the red arrows. **E.** Comparison of *E. coli* DH10β Δ *coaD* pTKcoaD cells grown at 30°C (left) and 42°C (right). Cells were grown in LB+Kan for 15 hours at the corresponding temperature, to allow for sufficient outgrowth, before plating on LB+Kan and incubating at the corresponding temperature. By comparing the number of GFP-positive colonies seen in each case we estimated an escape frequency of 1 in 16,500 (σ = 1,600). We also tracked the escape frequency of cells after transformation with PPAT homologs and growth at 42°C, by determining the ratio of GFP negative to GFP positive cells, finding an escape frequency of 1 in 20,200 (σ = 9500) as determined by 8 independent transformations. These escape frequencies are similar to those previously reported for *coaD* (a.k.a. *kdtB*) upon heat curing of *coaD* expressing pMAK705 plasmid in a conditional knockout [45].

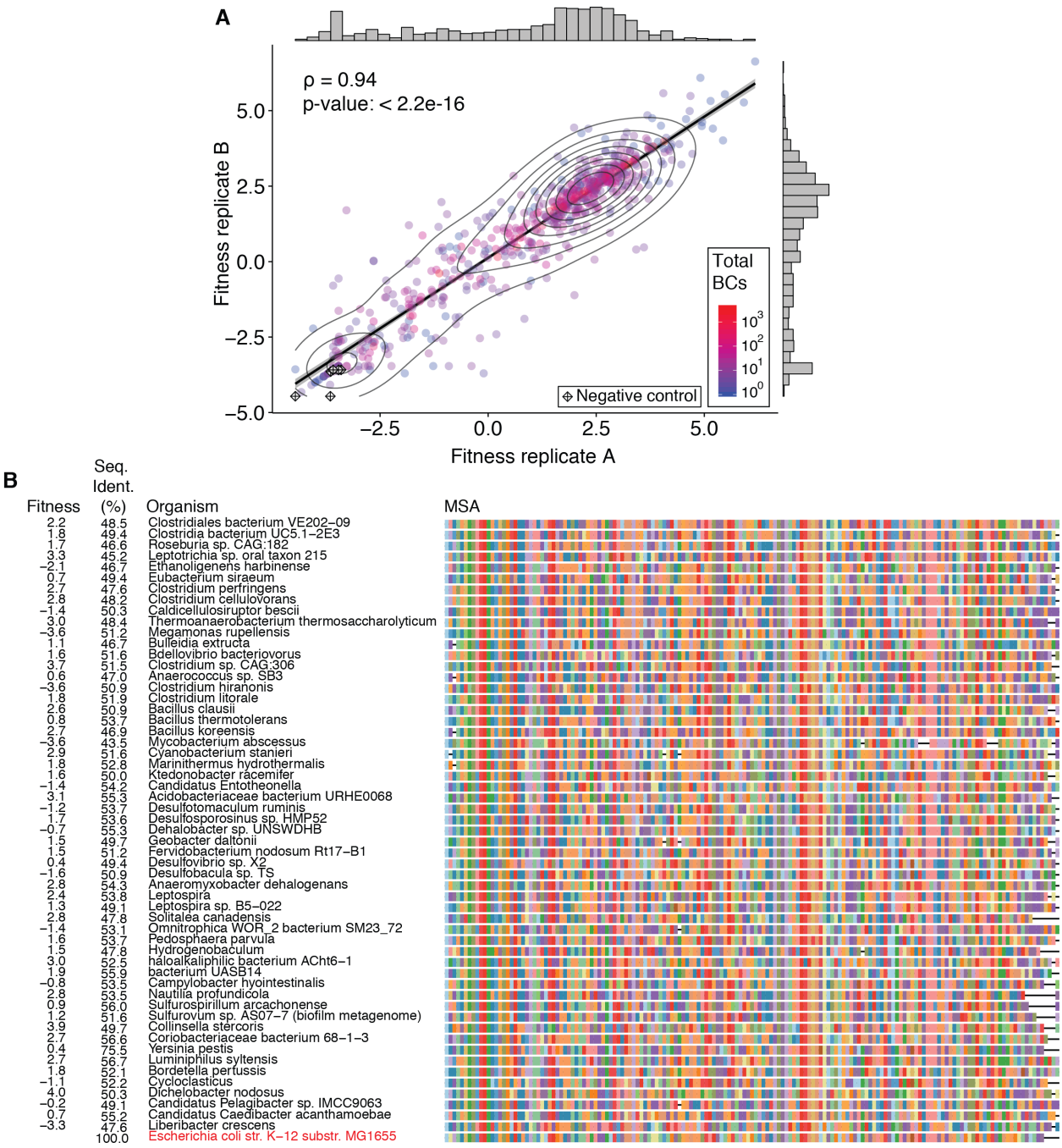

Figure 3.17: **PPAT complementation assay. A.** The fitness values for 651 homologs across two independent biological replicates shows strong correlation (ρ=0.94; Pearson). Six negative controls lacking the H/TxGH motif required for nucleophilic attack on the α phosphate of the ATP have very low fitness values (<3) in the assay. We colored each point based on the number of assembly barcodes that corresponded to errorless constructs, and find that reproducibility among replicates improves with increasing number of assembly barcodes (Fig. 3.18B). **C.** Despite having a median 50% sequence identity, distant homologs are typically still able to complement the function of the native *E. coli* PPAT (bottom row). This multiple sequence alignment table shows the fitness scores, percent sequence identity to *E. coli* PPAT, and source organism.

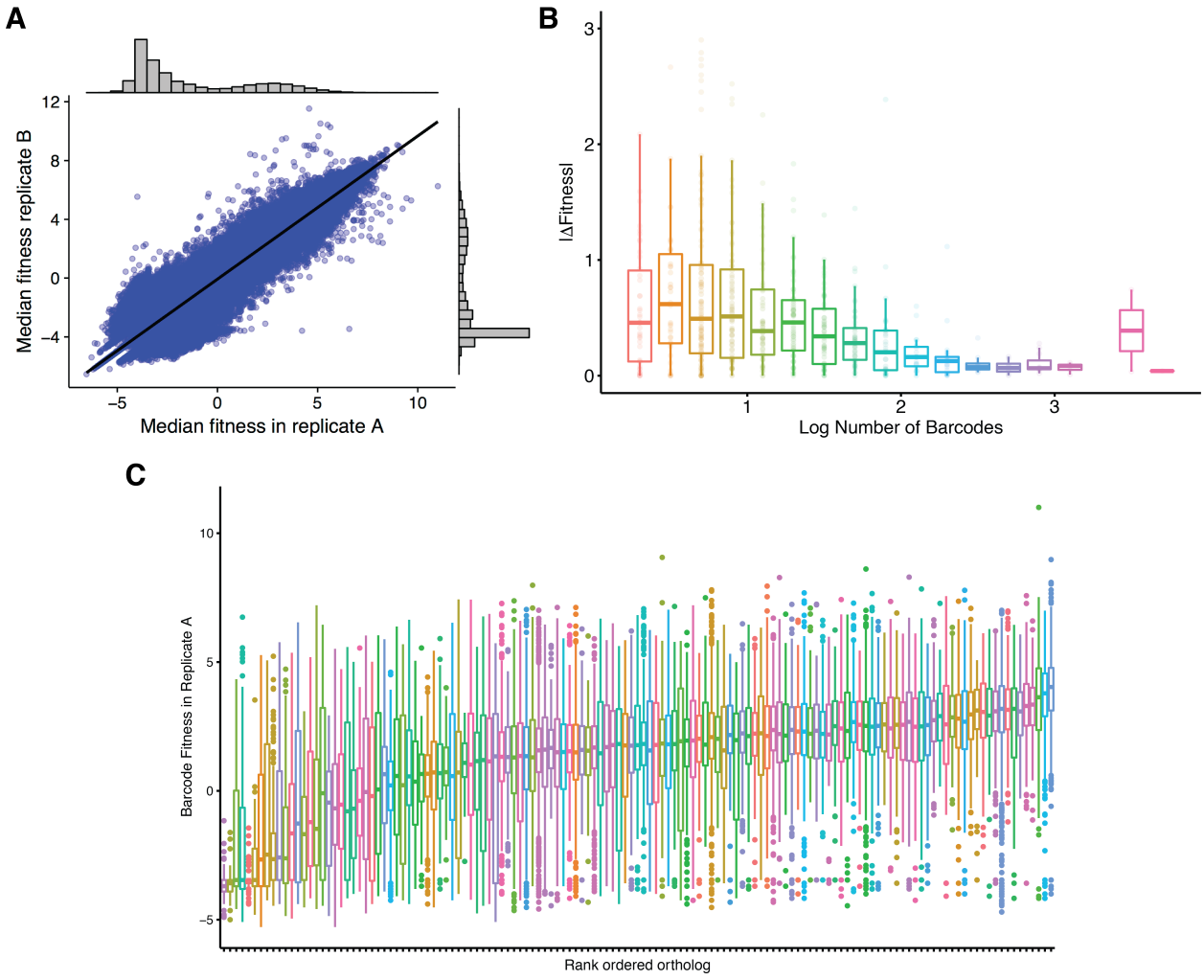

Figure 3.18: **A.** Fitness values of 329,897 individual assembly barcodes in each biological replicate, with a correlation of 0.948. A large number of low-fitness assembly barcodes correspond to assemblies with frameshifts due to indels. **B.** We see the reproducibility of the fitness values increase with the number of assembly barcodes. The absolute difference in homolog fitness values between the two biological replicates as a function of their number of assembly barcodes (ρ=-0.34; Spearman, p-value <2.2E-16). **C.** Fitness values are noisy with a median standard deviation of around 2.4. Box plots of individual assembly barcode fitness values for homologs in replicate A which have at least 50 assembly barcodes. Homologs are rank-ordered by their final fitness value.

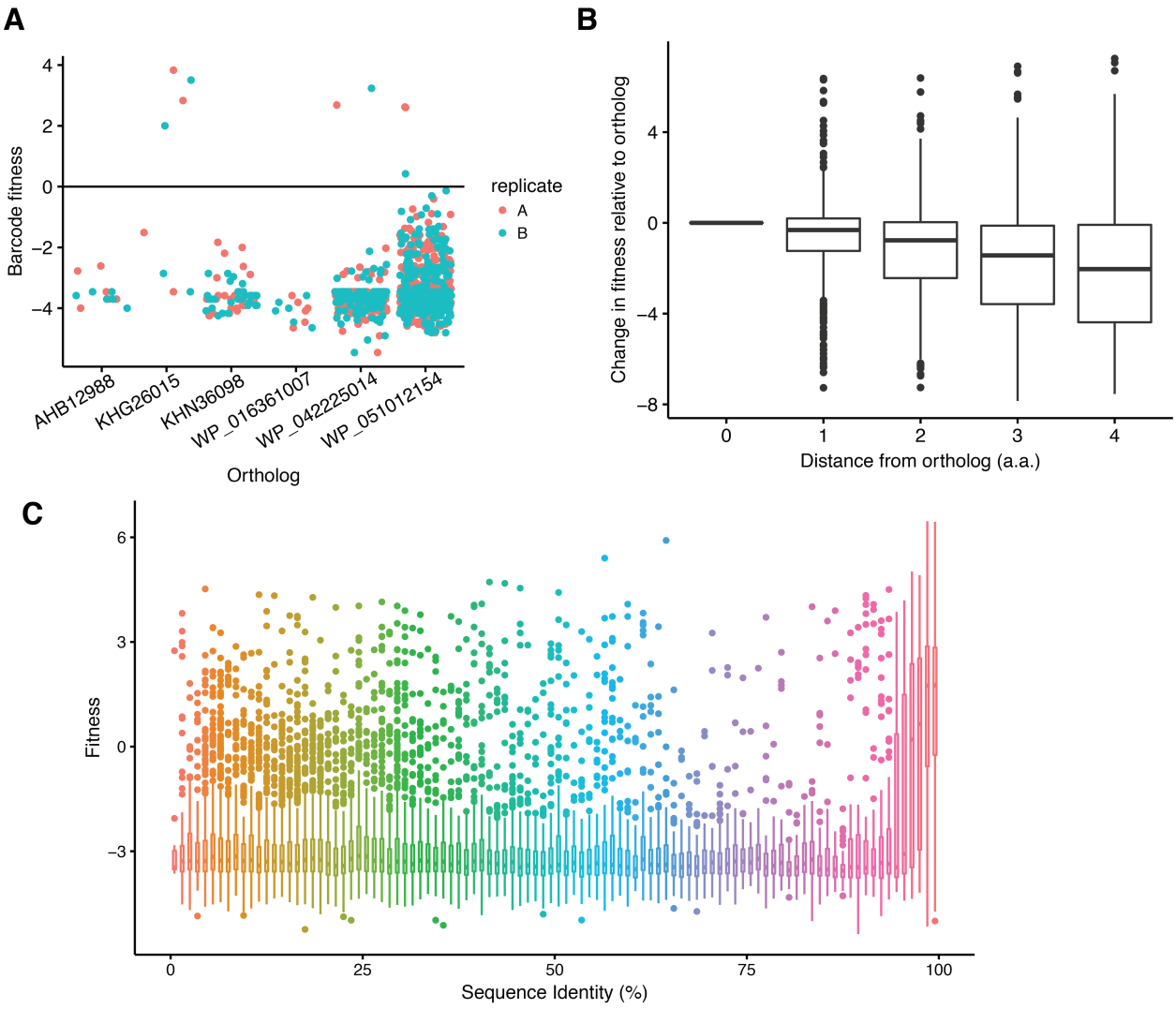

Figure 3.19: **A.** Assembly barcode fitness for six of the homologs missing the H/TxGH motif required for catalytic activity. No simple mutation would be able to restore catalytic activity to these homologs, so they serve as a useful measure of the false positive rate for individual assembly barcodes. Of the 994 assembly barcodes only 9 assembly barcodes (0.9%) have a positive fitness value, indicating a low rate of false positives at the individual barcode level. **B.** Mean sequence fitness is reduced with increasing number of mutations (ρ=-0.38; Spearman, p-value <2.2E-16). Analysis of 144,573 sequences' fitness as a function of their a.a. distance from the designed homolog sequence. **C.** Very few sequences with less than ~94% sequence identity show high fitness. For sequences represented by at least 2 assembly barcodes, we plot their fitness as a function of their sequence identity (relative to their corresponding designed sequences), within bins of 1%.

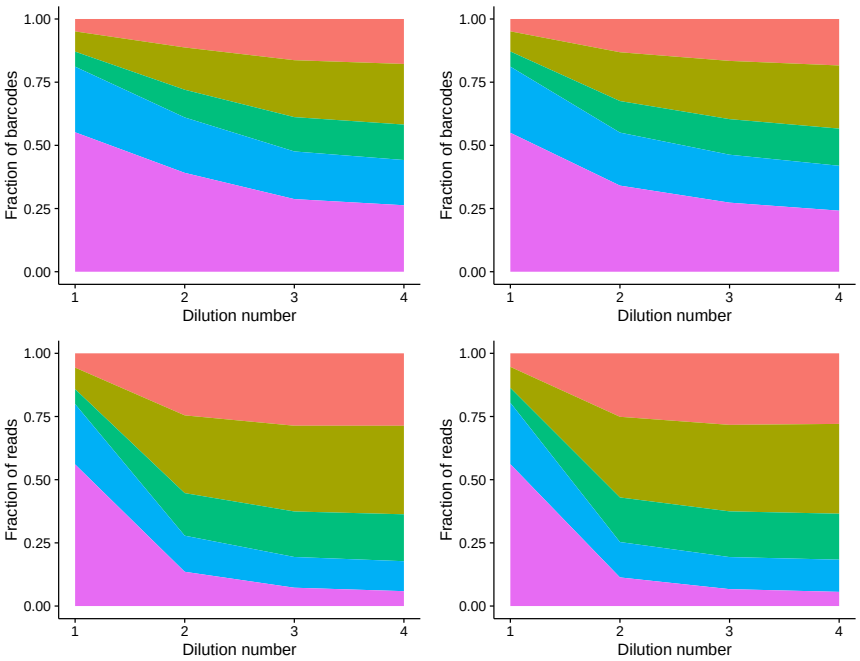

Figure 3.20: The population of perfect and low mutational distance sequences expand as a function of time, while sequences with low sequence identity (primarily due to indels) are depleted. We see that non-functional assemblies are lost from the population primarily between the first two dilutions. Distribution of mapped assembly barcodes **(top.** and mapped reads **(bottom.**, for each replicate **(left & right.**, based on distance from the designed sequence.

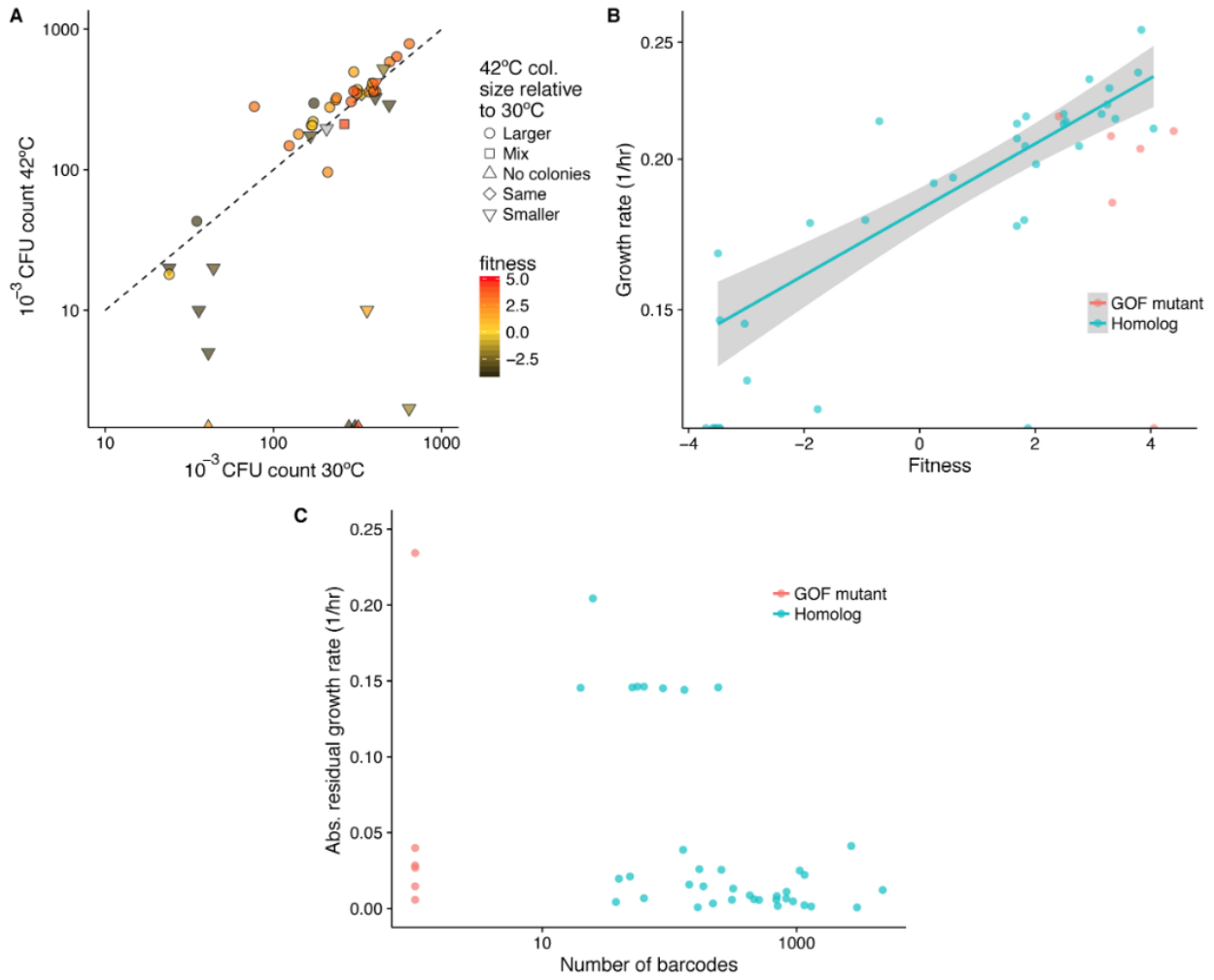

Figure 3.21: **Synthesis verification.** Sequence-verified clones were obtained for 37 of 49 homologs. **A.** The amount of colonies observed after transformation of amplified constructs into *E. coli* DH10β Δ *coaD* pTKcoaD cells grown at 30°C (positive control) and 42°C (complementation). Symbol indicates 42°C colony size relative to 30°C colonies. Dashed line shows slope of one and is not a fit. The presence of a cluster with low colony counts in both conditions made up primarily of low-fitness homologs suggests possible toxicity effects. Two false positives are observed which had positive fitness in the pooled assay but produced no colonies in this transformation. Both of these had a low number of assembly barcodes (1 and 25). The majority of high fitness homologs produced large numbers of colonies in both conditions with high correspondence between the two. **B.** Comparison of growth rate of individual homologs (log-scale) and gain-of-function mutants as determined on a plate reader with experimentally-determined fitness from pooled complementation assay, with a Spearman's correlation of r_S=0.86. Growth rate (hr^{-1}) is defined as the maximum slope of OD600 vs. time on a log/linear plot. Fit is carried out using log growth rate and does not include the eight homologs with a growth rate of zero. Wildtype PPAT E. coli had a growth rate of 0.132 indicative of gene dosage toxicity effects due to overexpression. **C.** Correlation between the residual error of the fit of growth rate to fitness and number of assembly barcodes in homologs (r_s=-0.50, Spearman, p-value 1.7E-3). Constructs with fewer assembly barcodes tend to have higher error between individual growth rate and fitness in the pooled assay, highlighting the need for many assembly barcodes to determine fitness.

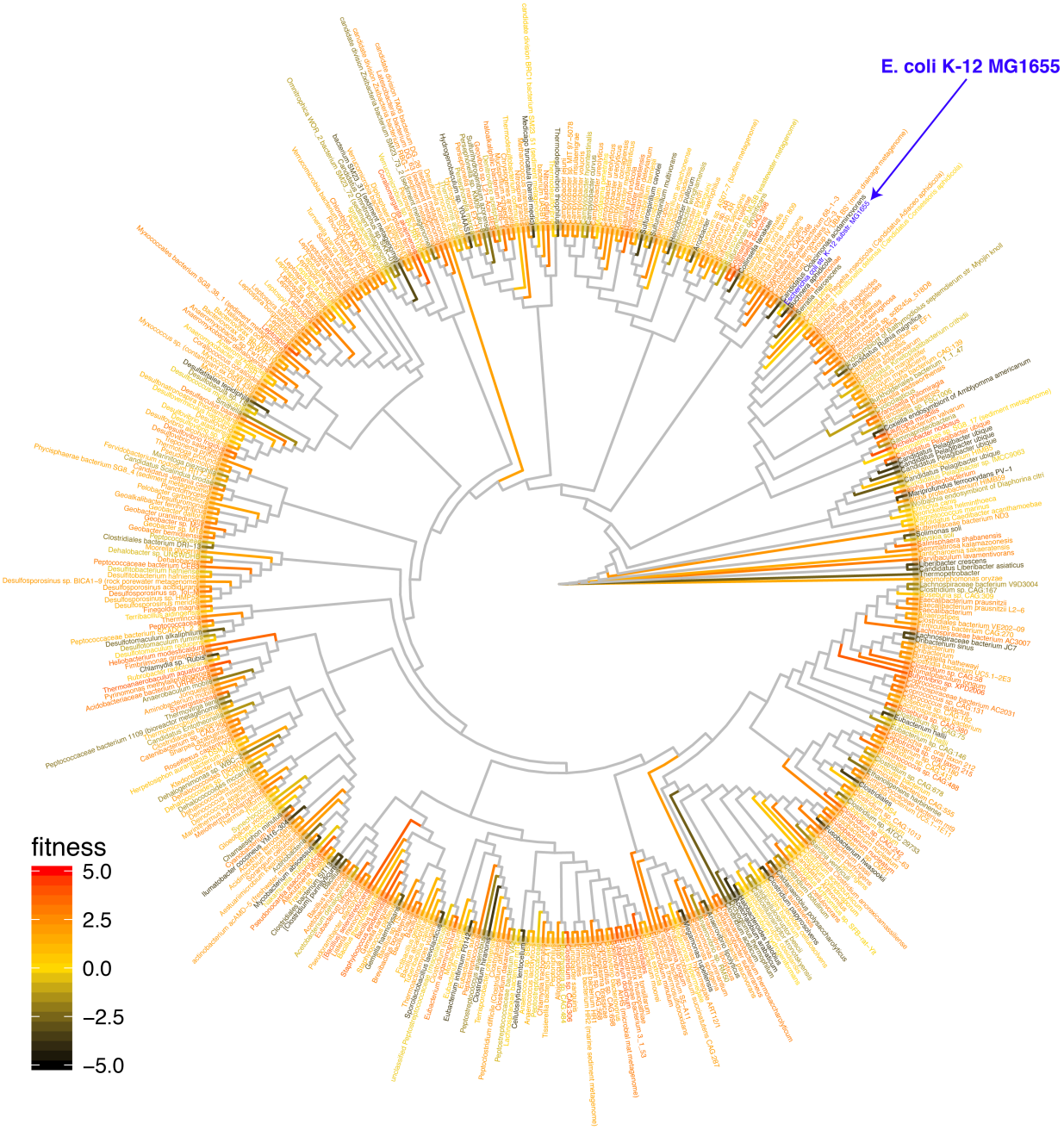

Figure 3.22: **PPAT phylogenetic tree.** The majority of homologs listed complement wildtype *E. coli*, with low-fitness homologs randomly dispersed throughout the tree with minimal clustering. A phylogenetic tree of 451 homologs labeled, similar to Fig. 3.3D, with each leaf labeled with the organism name and shaded by fitness.

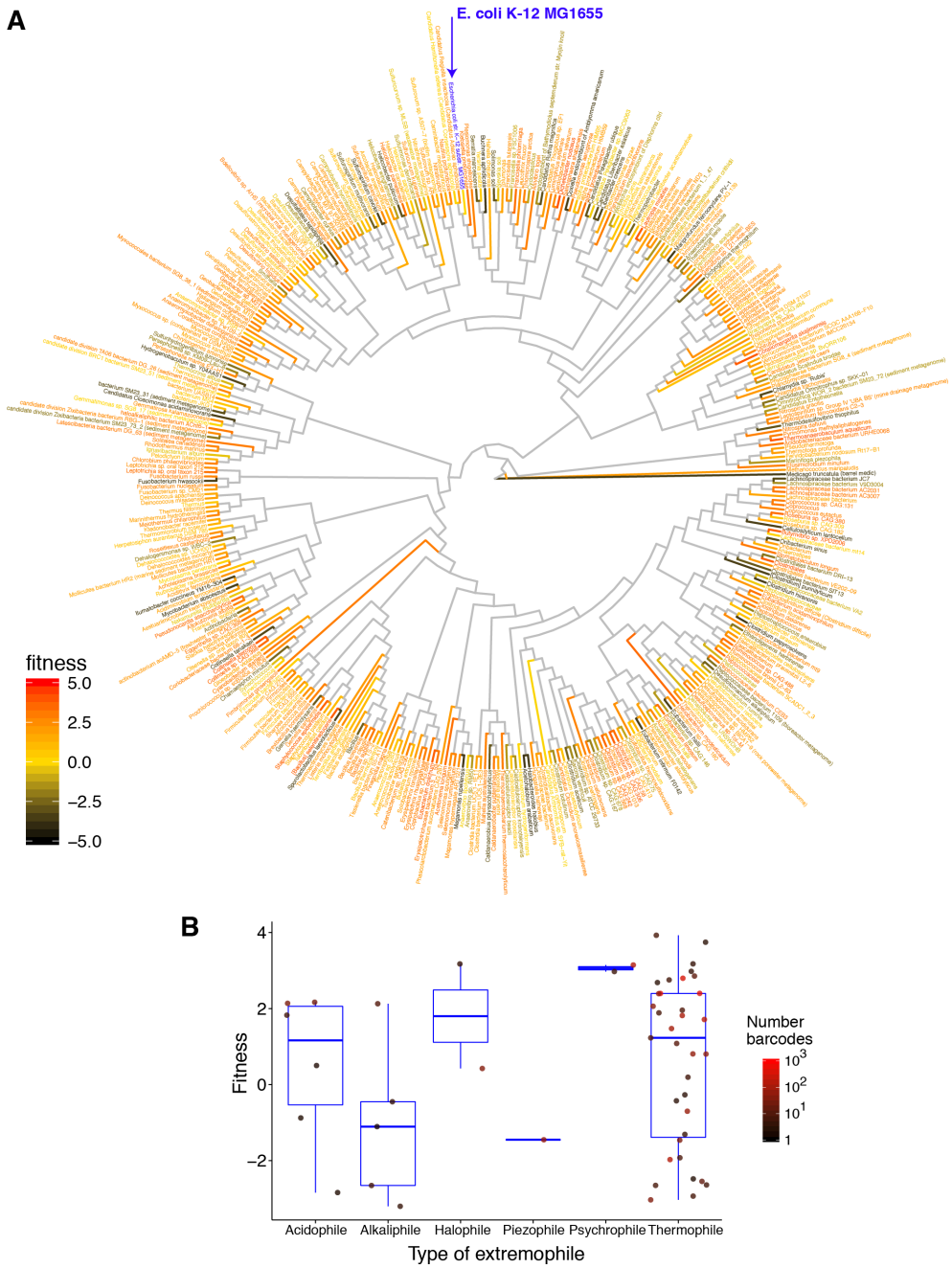

Figure 3.23: **A.** Phylogenetic tree of 411 homologs based on NCBI taxonomy rather than PPAT sequence, generated using phyloT (`http://phylot.biobyte.de`). The median fitness was used when multiple sequences were annotated with the same taxonomic ID. **B.** Fitness of PPAT homologs from organisms annotated as extremophiles. Of the different classes, alkaliphiles show a weak shift to lower fitness values (p=0.059 Wilcoxon rank sum test). Previous characterization of *E. coli* PPAT showed a maximum activity at pH 6.9 which was reduced to 68% of the maximum by pH 8 [48].

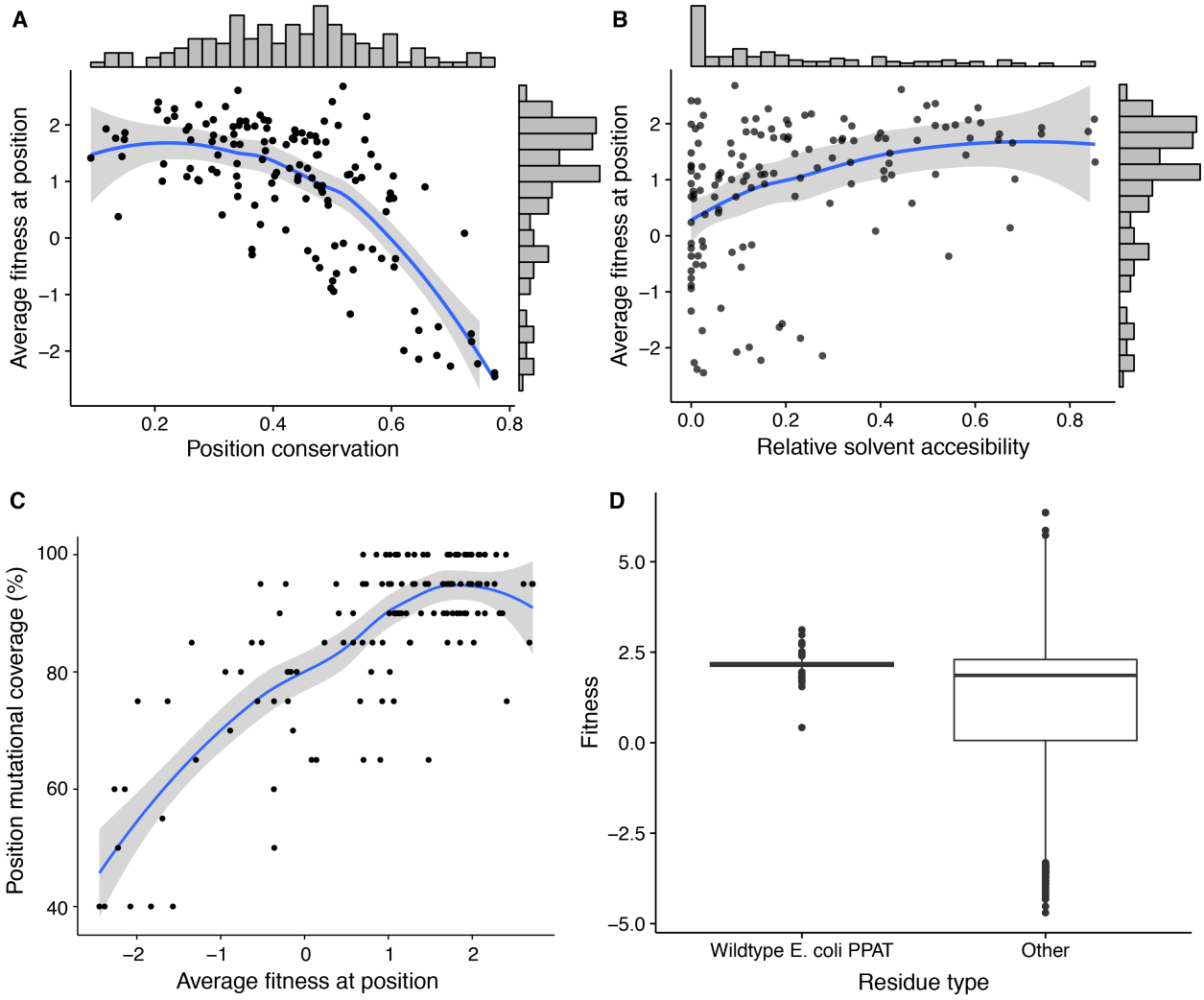

Figure 3.24: **A.** The average BMS position fitness compared to the conservation (Jensen-Shannon divergence). As expected mutations tend to be more constrained at highly conserved sites (ρ=-0.64; Pearson, p-value <2.2E-16). **B.** The average BMS position fitness compared to the relative solvent accessibility based on a DSSP analysis of the 1H1T crystal structure (dimer not hexamer). Buried residues tend to be more constrained (ρ=0.42; Pearson, p-value 3.9E-8). **C.** Mutational scanning coverage decreases at site of low fitness (ρ=0.76; Pearson, p-value <2.2E-16). This effect is due to assembly barcodes with low read numbers which, due to their low fitness, never pass the minimum 10 read threshold. **D.** Residues appearing in wildtype E. coli PPAT are associated with higher fitness values. The distribution of fitness values for residues present in the E. coli PPAT sequence (median = 2.16, $\sigma = 0.24$) compared to all others (median = 1.86, $\sigma = 2.16$).

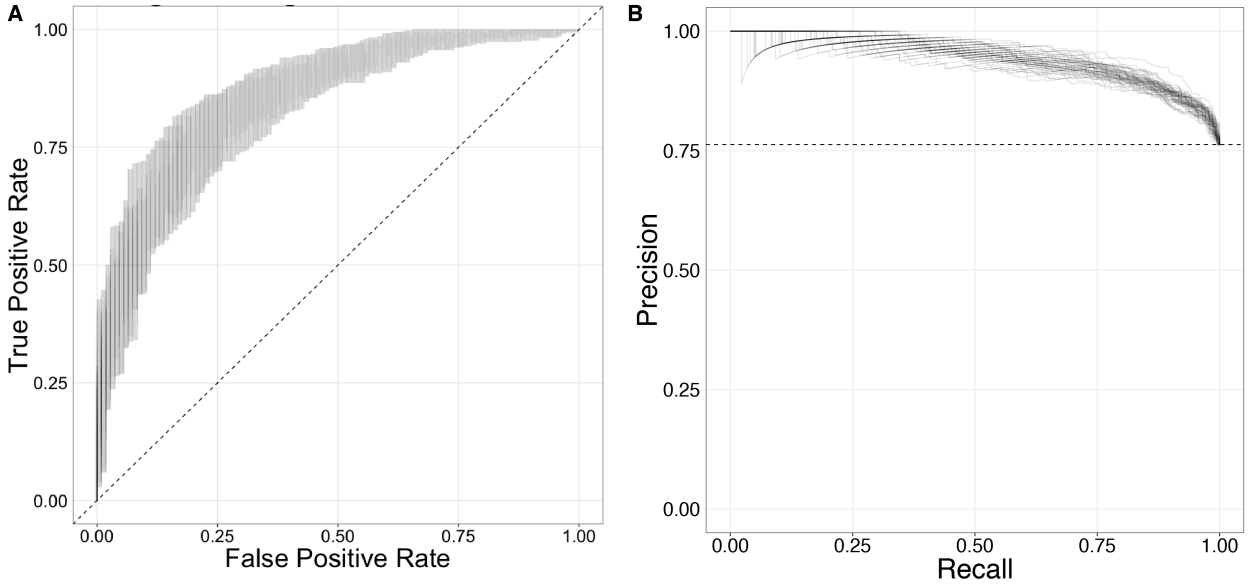

Figure 3.25: **Variant classifier.** We implemented a classifier to predict how different BMS variants would perform in our assay. Each BMS variant was categorized into two bins based on whether or not their measured fitness score was greater than 0. We then performed a logistic regression using 6 features for our model - the amino acid mutation, secondary structure class as assigned by DSSP (loop, beta-sheet, or alpha-helix), relative solvent accessibility as assigned by DSSP, sequence conservation, evolutionary coupling as predicted by EVMutation, and the frequency of residue substitution from the sequence alignment used for EVMutation's prediction. To assess the performance of our classifier, we performed 10 repeats of 5-fold cross-validation on our dataset and measured the precision and recall of each model on its respective hold-out set. We found that on average, our simple classifier has **A.** an average accuracy of 0.825 +/- 0.013, **B.** a precision of 0.853 +/- 0.009, and an average recall of 0.931 +/- 0.014.

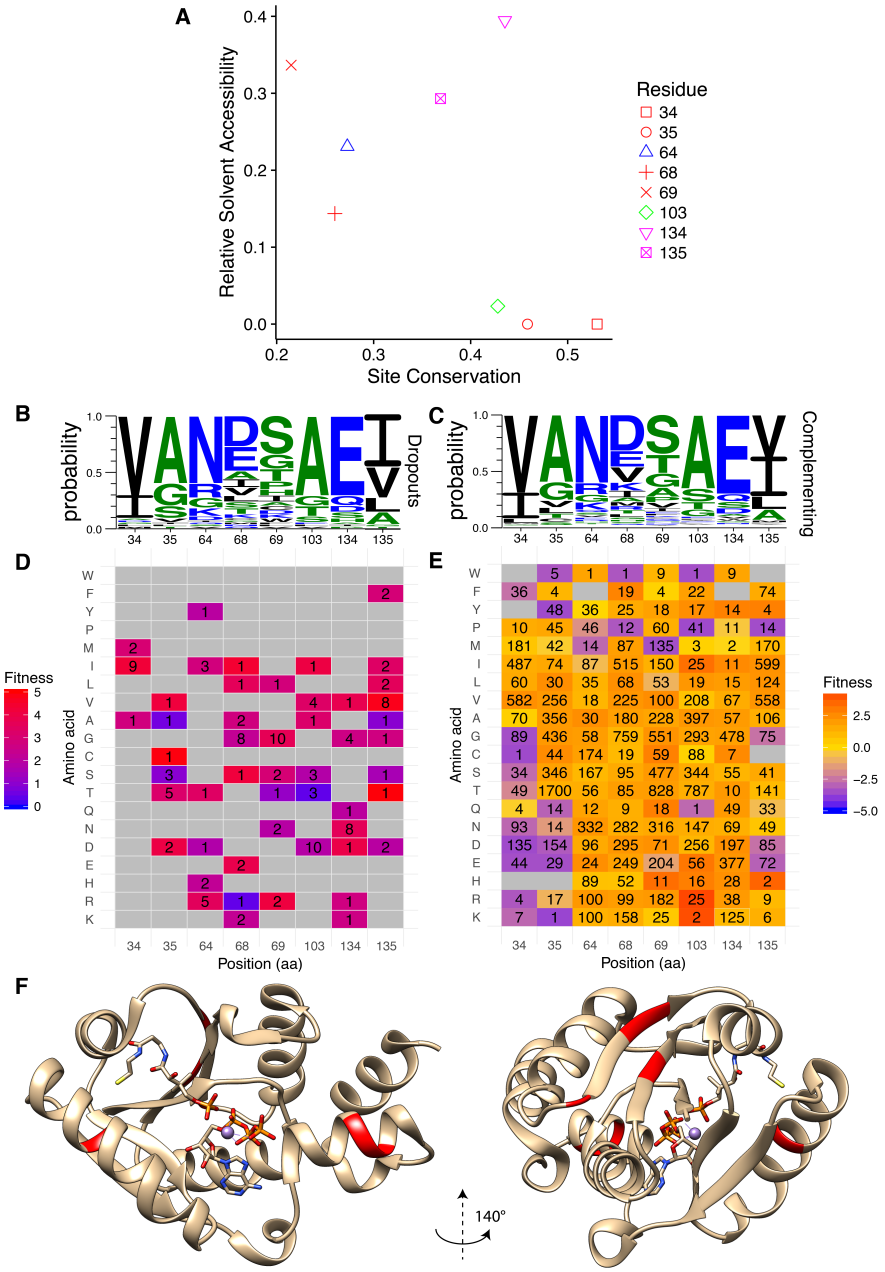

Figure 3.26: **A.** The relative solvent accessibility and conservation of each of the eight gain of function positions. **B.** Weblogo showing the probability of each residue at the gain-of-function positions for low-fitness homologs. **C.** Weblogo of GoF residues for homologs which complemented. **D.** The mean fitness of each GoF mutation at the significant positions, with the number of mutants observed at each a.a. **E.** The same plot with the data derived from the broad mutational scan using complementing homologs and their mutants. **F.** *E. coli* PPAT structure with the eight GoF residues shaded in red. Glu-134 is involved in hydrophobic interactions with coenzyme A [42], suggesting a role for GoF mutations in modulating the inhibitory feedback, while Ala-103 participates in hydrophobic interactions between the PPAT dimers.

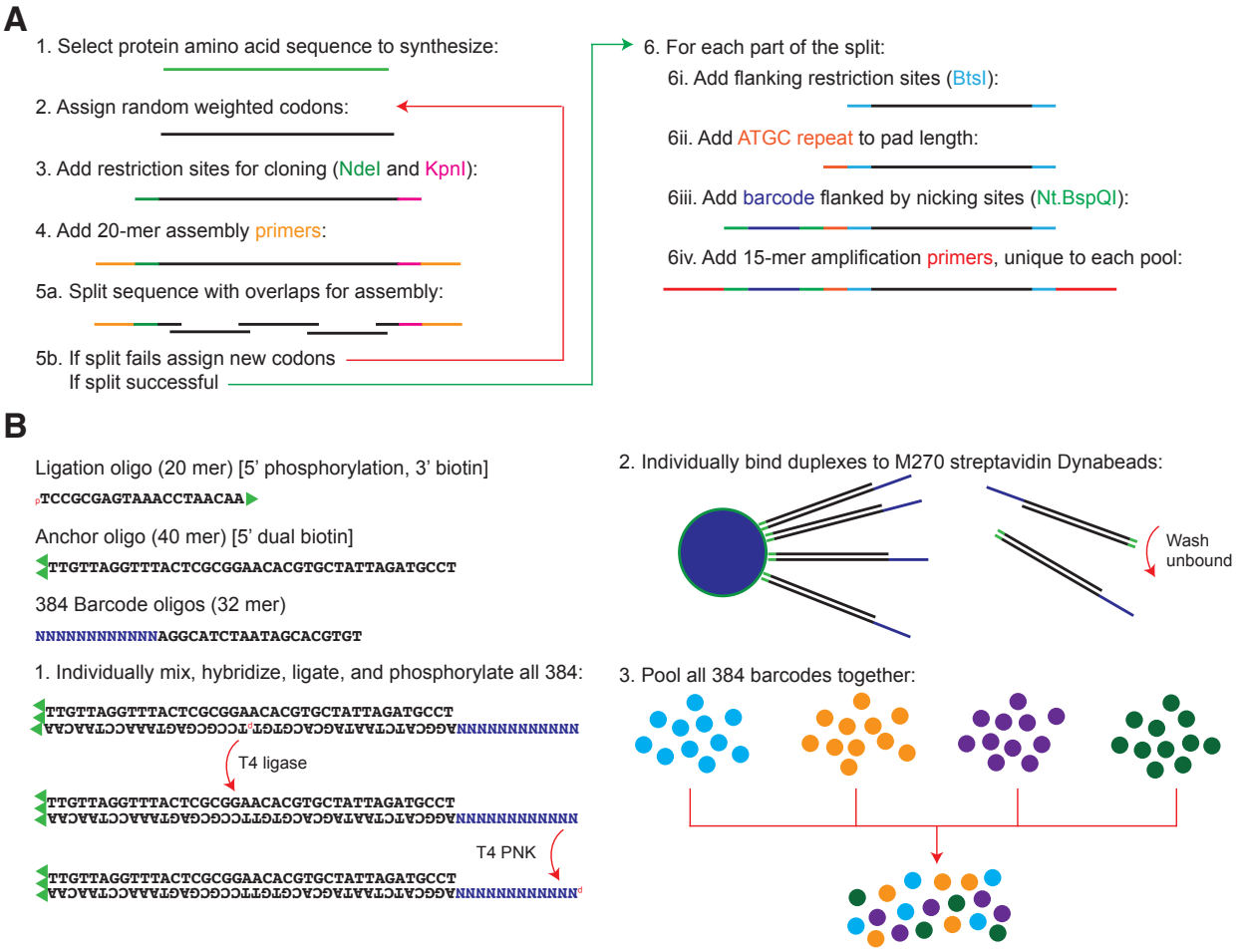

Figure 3.27: **A.** The oligo design process. Briefly, a.a sequences are assigned random weighted codons and appended with restriction and primer sites used in DropSynth assembly. Sequences are then split into five oligos with ∼20-nt overlap regions. Individual oligo sequences are appended with restriction sites, padding sequences, gene-specific microbead barcodes flanked by nicking sites, and amplification primer sites leading to a library of 200-nt sequences. **B.** The DropSynth microbead barcoding process. Microbead barcode oligos are individually mixed with 3' biotinylated ligation oligos and dual 5' biotinylated anchor oligos, ligated using T4 ligase and phosphorylated with T4 PNK, exposing the microbead barcode sequence (NNNNNNNNNNNN). Biotinylated duplexes are then individually bound to M270 streptavidin Dynabeads and pooled together.

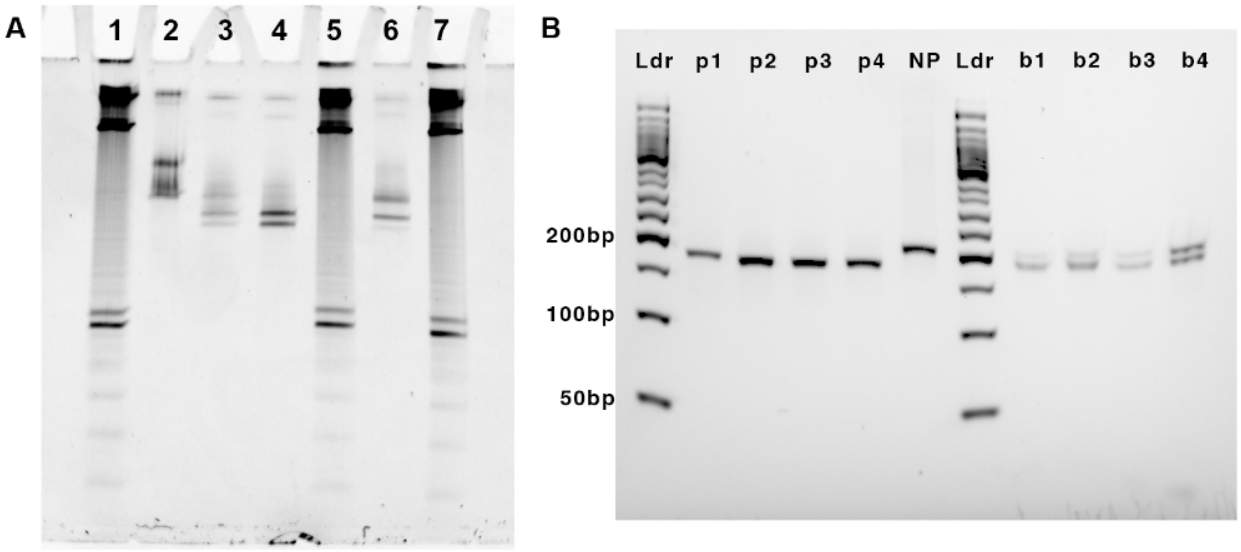

Figure 3.28: **Nick processing to generate single-stranded microbead barcode overhang. A.** A 10% TBE-Urea denaturing gel highlighting the steps in nick processing. Lanes 1, 5, 7: a 10 bp ladder. Lane 2: Before processing, all oligos should be 200 nt. Lane 3: After nick processing we expect fragments of 165 nt, 177 nt, 35 nt, and 23 nt. Lane 4: After streptavidin Dynabead cleanup of nick processed oligos we expect fragments of 165 nt and 177 nt. Lane 6: The captured Dynabead fraction after boiling at 90°C for 10 min in 10 mM EDTA pH 8.2. **B.** A non-denaturing 4% agarose gel showing the nick processing which takes a 200 bp duplex and leaves a 12-nt single-stranded microbead barcode overhang on a 165 bp dsDNA fragment. Lanes p1-p4 showing several samples after nick processing and also one before processing (NP). Lanes b1-b4 show the corresponding Dynabead fractions after denaturing at 80°C for 3 min. Full length oligos containing errors in the nt.BspQI sites will not have both strands nicked and are likely to be pulled down by the Dynabeads together with the short fragment.

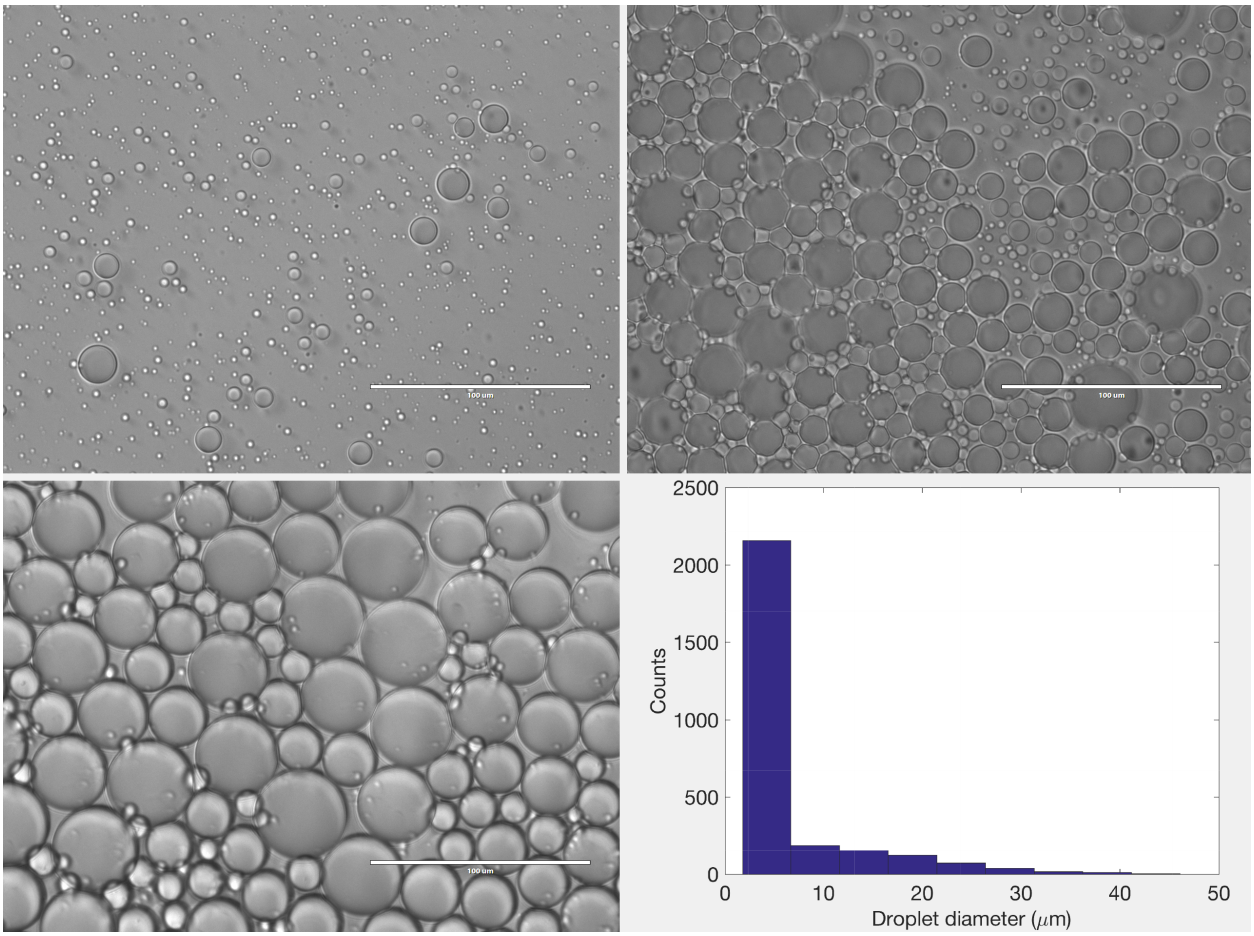

Figure 3.29: **Characterization of the distribution of droplet sizes for the vortex emulsions.** Briefly, 100 uL of Kapa Robust buffer was added to an eppendorf tube with 600 uL of Bio-Rad Droplet Generation Oil and vortexed upright for 4 minutes on the highest setting of a Vortex-Genie 2. Samples were then taken from the bottom, middle, and top of the resulting emulsion and imaged under 40X magnification. The mode of the droplet diameter distribution peaks below 5 um. Scale bars are 100 um. Bottom right: Histogram of droplet diameters as determined by image analysis. Median droplet diameter is below 5 um.

Table 3.1: **Assembly barcode statistics for each serial dilution in the two biological replicates.** Barcodes for each sample were clustered using Starcode [35] to collapse barcodes within a Levenshtein distance of 1.

Biological replicates	**Serial dilution**	**Total reads**	**Total baracodes**	**Total clutered barcodes**
A	1	9,051,752	4,317,940	4,289,165
	2	9,790,924	2,319,457	2,231,361
	3	8,222,783	1,346,284	1,263,430
	4	7,947,874	970,291	892,753
B	1	9,136,919	4,259,319	4,228,531
	2	8,319,364	1,919,591	1,843,449
	3	10,036,601	1,393,886	1,292,371
	4	9,437,037	993,877	907,884

Table 3.2: **Homologs and GoF mutants retrieved from the assembled library and individually tested in knockout (KO) PPAT cells.** Growth rate (hr^{-1}) is defined as the maximum slope of OD600 vs. time on a log/linear plot. Wildtype *E. coli* PPAT and 3 catalytically inactive wildtype mutants were also prepared and tested.

Type	**Construct ID**	**Assembly barcodes**	**42°C CFU**	**30°C CFU**	**Growth rate (hr^{-1}.**	**Fitness**
Homolog	CDD12392	699	312	234	0.198	2.02
Homolog	WP_041531153	63	343	334	0.178	-0.94
Homolog	CDA36762	834	324	237	0.205	2.76
Homolog	WP_051012154	63	0	281	0	-3.46
Homolog	WP_012984121	1302	372	315	0.205	1.83
Homolog	WP_028874703	462	357	371	0.208	1.69
Homolog	WP_009532117	220	349	313	0.218	3.15
Homolog	CDC50010	429	361	406	0.216	3.39
Homolog	WP_025936372	1150	207	168	0.215	2.53
Homolog	WP_050330521	89	20	24	0	-3.58
Homolog	WP_028844278	38	174	166	0.146	-3.03
Homolog	WP_012096847	710	383	383	0.214	2.49
Homolog	WP_050708028	1154	362	300	0.212	4.05
Homolog	WP_027397238	40	20	44	0.131	-2.99
Homolog	WP_007413164	172	497	300	0.176	1.68
Homolog	KOS35328	25	0	41	0	1.88
Homolog	KHS64893	506	96	210	0.218	2.49
Homolog	KGB86419	185	585	491	0.233	2.94
Homolog	WP_025369197	131	0	305	0	-3.70
Homolog	WP_021271192	256	10	360	0.178	1.81
Homolog	KJF18279	242	10	36	0	-3.52

Homolog	WP_029522041	128	2	640	0.124	-1.77
Homolog	WP_014806494	143	523	451	0.177	-1.90
Homolog	WP_038558636	938	18	24	0.193	0.58
Homolog	CDC19414	310	638	541	0.236	3.78
Homolog	WP_029455214	49	43	35	0.167	-3.49
Homolog	BAN02173	56	290	487	0	-3.46
Homolog	WP_008711239	693	787	642	0.229	3.28
Homolog	WP_039669974	51	5	41	0	-3.52
Homolog	WP_013656808	167	326	403	0.147	-3.46
Homolog	KJS87341	2691	220	172	0.215	-0.70
Homolog	WP_005674855	1059	305	290	0.256	3.84
Homolog	WP_011140849	4757	278	216	0.214	1.68
Homolog	WP_009360218	2986	281	77	0.222	3.25
Homolog	EUC78355	317	413	388	0.217	1.84
Homolog	WP_011433776	828	206	170	0.191	0.24
Homolog	WP_006440043	20	297	174	0	-3.55
GOF	WP_013656808_S69R	1	148	124	0.209	3.31
GOF	WP_029455214_K69T	1	361	394	0.184	3.34
GOF	WP_023508997_A104V	1	179	141	0.217	2.41
GOF	WP_049662705_A101V	1	415	404	0.204	3.82
GOF	WP_054252071_D66E	1	0	319	0	4.06
GOF	WP_044825986_V134F	1	211	264	0.211	4.40
Wildtype	NP_418091		196	207	0.132	ND
Inactive	NP_418091_H18Y		95	142	0.105	
Inactive	NP_418091_H18D		0	130	0	
Inactive	NP_418091_H18W		0	113	0	

Table 3.3: **Cost to create pool of 384 barcoded DropSynth microbeads.** Creating the pool of barcoded beads is a one time cost and produces enough beads to carry out at least 210 assemblies of 384 genes, or over 80,000 genes, using the current protocol.

Item	**Cost**
38.4 nmol anchor oligo (5' dual biotin modification)	$300
38.4 nmol ligation oligo (5' phosphorylation and 3' biotin modifications)	$540
0.1 nmol of each of the 384 barcoded oligos	$1656
1,575 uL 10X T4 ligase buffer	$5
80E9 Units of T4 ligase (concentrated)	$40
1,920 uL stock M270 streptavidin Dynabeads	$456
3.84E9 Units T4 PNK	$344
TOTAL	**$3341**
Cost per assembly	**$15.69**
Cost per construct	**$0.04**

Table 3.4: **DropSynth assembly costs per 384 gene library.**

Item	**Cost**
Microarray derived OLS pool (Agilent Technologies; ∼$0.10/oligo)	$192
3x 50uL rxn KAPA Real-time Library Amplification Kit (KAPA Biosystems)	$8.4
8x 50uL rxn KAPA HiFi HotStart ReadyMix (KAPA Biosystems)	$33.6
2x columns Zymo Clean & Concentrator -5 (Zymo Research)	$2.5
1 biotinylated primer (Integrated DNA Technologies)	$40
3 primers (Integrated DNA Technologies)	$10
2x SPRI cleanup (Beckman)	$1
1,200 uL Dynabeads M-270 Streptavidin (Invitrogen)	$285
50 uL Nicking enzyme (Nt.BspQI) (New England Biolabs)	$27
50 uL Kapa2G Robust HotStart ReadyMix (KAPA Biosystems)	$2.7
7 uL BtsI (New England Biolabs)	$7.3
600 uL BioRad Droplet Generation Oil for EvaGreen (Bio-Rad Laboratories)	$2.3
TOTAL	**$612**
Cost per Construct	**$1.59**
Cost per Construct with Barcoded Beads	**$1.63**

Table 3.5: **Nick processing efficiencies for various conditions.**

Sample	nt.BspQI digest (min)	[nt.BspQI] (U/uL) in 150 uL digest	[DNA] (ng/uL) in digest	M270 Dynabead incubation (min)	[M270 Dynabead] (ug/uL)	Molar Yield (%)
i	190	0.266	9.5	200	2.2	56%
ii	240	0.266	14.4	280	1.2	35%
iii	985	0.3	16	375	2.2	46%
iv	985	0.32	14.4	375	1.2	40%
v	390	0.3	15.7	35	2.2	47%
vi	390	0.34	18.7	50	2.2	44%
vii	390	0.32	17.6	70	2.2	36%

Table 3.6: **The oligos required for the bead barcoding process.** All oligos were ordered from Integrated DNA Technologies.

Oligo Name	**Sequence**	**Modifications**	**Amount**
Ligation oligo	TCCGCGAGTAAACCTAACAA	3' biotin 5' phosphorylation	38.4 nmol
Anchor oligo	TTGTTAGGTTTACTCGCGGAA-CACGTGCTATTAGATGCCT	5' dual biotin	38.4 nmol
384X barcoded oligos	12-mer microbead barcode reverse complement + AGGCATCTAATAGCACGTGT		0.1 nmol each

Table 3.7: **Primer sequences used in this study.**

Name	Sequence
coaD_KO_KAN_FWD_1	AACGCATTGAGGTTGTTGAAGTTCCTATACTTTCTAGAGAATAGGAACTTCGGAATAGGAACTTCTTTCTTAGACGTCGGAATTGCCAGC
coaD_KO_KAN_REV_1	ATACCATCCGGCATAAACGAGTTCCTATTCCGAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCGCTCAGAAGAACTCGTCAAGAAGGC
coaD_KO_KAN_FWD_2	GCTTCAACTGCTGGAACCTTACCTGCCACCGAAAACGCATTGAGGTTGTTGAAGTTCC
coaD_KO_KAN_REV_2	TGCCAGAAGTAATTCATGCGCGCCGGATGGCATACCATCCGGCATAAACGAGTTCC
pEVBC_FWD	Biotin-GCCGTCATATGAGCTGTTTCCTGTGTGAAATTG
pEVBC_REV1	Biotin-GTGGGTACCTAAGTGTGGCTGCGGAACNNNNNNNNNNNNNNNNNNNNGCACGACGTCAGGTGGCACTTTTCG
pEVBC_amp_FWD	Biotin-GTGGGTACCTAAGTGTGGCTGCGGAAC
mi3_FWD	AATGATACGGCGACCACCGAGATCTACACGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGCTCATATG
mi3_REV_N70#	CAAGCAGAAGACGGCATACGAGATNNNNNNNNCGCACATTTCCCCGAAAAGTGCCACCTGACG
mi3_R1	GTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGCTCATATG
mi3_R2	CGCACATTTCCCCGAAAAGTGCCACCTGACGTCGTGC
mi3_index	GCACGACGTCAGGTGGCACTTTTCGGGGAAATGTGCG
mi4_FWD	AATGATACGGCGACCACCGAGATCTACACGCTAGACGGTACCTAAGTGTGGCTGCGGAAC
mi4_REV_N7##	CAAGCAGAAGACGGCATACGAGATNNNNNNNNCCTGACGTCAGGCAAGTGCCACCTGACGTCGTGC
mi4_R1	GGCTAGACGGTACCTAAGTGTGGCTGCGGAAC
mi4_R2	CCTGACGTCAGGCAAGTGCCACCTGACGTCGTGC
mi4_index	GCACGACGTCAGGTGGCACTTGCCTGACGTCAGG

References

[1] F. Inoue and N. Ahituv, "Decoding enhancers using massively parallel reporter assays," *Genomics*, vol. 106, pp. 159–164, Sept. 2015.

[2] M. Gasperini, L. Starita, and J. Shendure, "The power of multiplexed functional analysis of genetic variants," *Nat. Protoc.*, vol. 11, pp. 1782–1787, Oct. 2016.

[3] K. S. Sarkisyan, D. A. Bolotin, M. V. Meer, D. R. Usmanova, A. S. Mishin, G. V. Sharonov, D. N. Ivankov, N. G. Bozhanova, M. S. Baranov, O. Soylemez, N. S. Bogatyreva, P. K. Vlasov, E. S. Egorov, M. D. Logacheva, A. S. Kondrashov, D. M. Chudakov, E. V. Putintseva, I. Z. Mamedov, D. S. Tawfik, K. A. Lukyanov, and F. A. Kondrashov, "Local fitness landscape of the green fluorescent protein," *Nature*, vol. 533, pp. 397–401, May 2016.

[4] D. M. Fowler and S. Fields, "Deep mutational scanning: a new style of protein science," *Nat. Methods*, vol. 11, pp. 801–807, Aug. 2014.

[5] G. J. Rocklin, T. M. Chidyausiku, I. Goreshnik, A. Ford, S. Houliston, A. Lemak, L. Carter, R. Ravichandran, V. K. Mulligan, A. Chevalier, C. H. Arrowsmith, and D. Baker, "Global analysis of protein folding using massively parallel design, synthesis, and testing," *Science*, vol. 357, pp. 168–175, July 2017.

[6] S. Kosuri and G. M. Church, "Large-scale de novo DNA synthesis: technologies and applications," *Nat. Methods*, vol. 11, pp. 499–507, May 2014.

[7] S. Ma, N. Tang, and J. Tian, "DNA synthesis, assembly and applications in synthetic biology," *Curr. Opin. Chem. Biol.*, vol. 16, pp. 260–267, Aug. 2012.

[8] J. Quan, I. Saaem, N. Tang, S. Ma, N. Negre, H. Gong, K. P. White, and J. Tian, "Parallel on-chip gene synthesis and application to optimization of protein expression," *Nat. Biotechnol.*, vol. 29, pp. 449–452, May 2011.

[9] R. A. Hughes and A. D. Ellington, "Synthetic DNA synthesis and assembly: Putting the synthetic in synthetic biology," *Cold Spring Harb. Perspect. Biol.*, vol. 9, Jan. 2017.

[10] S. Kosuri, N. Eroshenko, E. M. Leproust, M. Super, J. Way, J. B. Li, and G. M. Church, "Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips," *Nat. Biotechnol.*, vol. 28, pp. 1295–1299, Dec. 2010.

[11] A. Y. Borovkov, A. V. Loskutov, M. D. Robida, K. M. Day, J. A. Cano, T. Le Olson, H. Patel, K. Brown, P. D. Hunter, and K. F. Sykes, "High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides," *Nucleic Acids Res.*, vol. 38, p. e180, Oct. 2010.

[12] J. C. Klein, M. J. Lajoie, J. J. Schwartz, E.-M. Strauch, J. Nelson, D. Baker, and J. Shendure, "Multiplex pairwise assembly of array-derived DNA oligonucleotides," *Nucleic Acids Res.*, vol. 44, p. e43, Mar. 2016.

[13] H. Kim, H. Han, J. Ahn, J. Lee, N. Cho, H. Jang, H. Kim, S. Kwon, and D. Bang, "'shotgun DNA synthesis' for the high-throughput construction of large DNA molecules," *Nucleic Acids Res.*, vol. 40, p. e140, Oct. 2012.

[14] T. H.-C. Hsiau, D. Sukovich, P. Elms, R. N. Prince, T. Strittmatter, T. Stritmatter, P. Ruan, B. Curry, P. Anderson, J. Sampson, and J. C. Anderson, "A method for multiplex gene synthesis employing error correction based on expression," *PLoS One*, vol. 10, p. e0119927, Mar. 2015.

[15] J. J. Schwartz, C. Lee, and J. Shendure, "Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules," *Nat. Methods*, vol. 9, pp. 913–915, Sept. 2012.

[16] N. B. Lubock, D. Zhang, A. M. Sidore, G. M. Church, and S. Kosuri, "A systematic comparison of error correction enzymes by next-generation sequencing," *Nucleic acids research*, vol. 45, no. 15, pp. 9206–9217, 2017.

[17] T. Izard and A. Geerlof, "The crystal structure of a novel bacterial adenylyltransferase reveals half of sites reactivity," *EMBO J.*, vol. 18, pp. 2021–2030, Apr. 1999.

[18] B. L. M. de Jonge, G. K. Walkup, S. D. Lahiri, H. Huynh, G. Neckermann, L. Utley, T. J. Nash, J. Brock, M. San Martin, A. Kutschke, M. Johnstone, V. Laganas, L. Hajec, R.-F. Gu, H. Ni, B. Chen, K. Hutchings, E. Holt, D. McKinney, N. Gao, S. Livchak, and J. Thresher, "Discovery of inhibitors of 4'-phosphopantetheine adenylyltransferase (PPAT) to validate PPAT as a target for antibacterial therapy," *Antimicrob. Agents Chemother.*, vol. 57, pp. 6005–6015, Dec. 2013.

[19] S. Bhattacharyya, S. Bershtein, J. Yan, T. Argun, A. I. Gilson, S. A. Trauger, and E. I. Shakhnovich, "Transient protein-protein interactions perturb e. coli metabolome and cause gene dosage toxicity," *Elife*, vol. 5, p. e20309, 2016.

[20] T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. I. Schärfe, M. Springer, C. Sander, and D. S. Marks, "Mutation effects predicted from sequence co-variation," *Nat. Biotechnol.*, vol. 35, pp. 128–135, Feb. 2017.

[21] T. Izard, "The crystal structures of phosphopantetheine adenylyltransferase with bound substrates reveal the enzyme's catalytic mechanism," *J. Mol. Biol.*, vol. 315, pp. 487–495, Jan. 2002.

[22] D. S. Marks, T. A. Hopf, and C. Sander, "Protein structure prediction from sequence variation," *Nat. Biotechnol.*, vol. 30, pp. 1072–1080, Nov. 2012.

[23] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, "Protein sectors: evolutionary units of three-dimensional structure," *Cell*, vol. 138, pp. 774–786, Aug. 2009.

[24] C. Notredame, D. G. Higgins, and J. Heringa, "T-Coffee: A novel method for fast and accurate multiple sequence alignment," *J. Mol. Biol.*, vol. 302, pp. 205–217, Sept. 2000.

[25] A. Stamatakis, "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies," *Bioinformatics*, vol. 30, pp. 1312–1313, May 2014.

[26] R. D. Finn, T. K. Attwood, P. C. Babbitt, A. Bateman, P. Bork, A. J. Bridge, H.-Y. Chang, Z. Dosztányi, S. El-Gebali, M. Fraser, J. Gough, D. Haft, G. L. Holliday, H. Huang, X. Huang, I. Letunic, R. Lopez, S. Lu, A. Marchler-Bauer, H. Mi, J. Mistry, D. A. Natale, M. Necci, G. Nuka, C. A. Orengo, Y. Park, S. Pesseat, D. Piovesan, S. C. Potter, N. D. Rawlings, N. Redaschi, L. Richardson, C. Rivoire, A. Sangrador-Vegas, C. Sigrist, I. Sillitoe, B. Smithers, S. Squizzato, G. Sutton, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, I. Xenarios, L.-S. Yeh, S.-Y. Young, and A. L. Mitchell, "InterPro in 2017-beyond protein family and domain annotations," *Nucleic Acids Res.*, vol. 45, pp. D190–D199, Jan. 2017.

[27] N. Eroshenko, S. Kosuri, A. H. Marblestone, N. Conway, and G. M. Church, "Gene assembly from Chip-Synthesized oligonucleotides," in *Current Protocols in Chemical Biology*, John Wiley & Sons, Inc., 2009.

[28] T. Buschmann and L. V. Bystrykh, "Levenshtein error-correcting barcodes for multiplexed DNA sequencing," *BMC Bioinformatics*, vol. 14, p. 272, Sept. 2013.

[29] W. J. Kent, "BLAT—The BLAST-Like alignment tool," *Genome Res.*, vol. 12, pp. 656–664, Apr. 2002.

[30] T. E. Kuhlman and E. C. Cox, "Site-specific chromosomal integration of large synthetic constructs," *Nucleic Acids Res.*, vol. 38, p. e92, Apr. 2010.

[31] C. Lou, B. Stanton, Y.-J. Chen, B. Munsky, and C. A. Voigt, "Ribozyme-based insulator parts buffer synthetic circuits from genetic context," *Nat. Biotechnol.*, vol. 30, pp. 1137–1142, Nov. 2012.

[32] J.-D. Pédelacq, S. Cabantous, T. Tran, T. C. Terwilliger, and G. S. Waldo, "Engineering and characterization of a superfolder green fluorescent protein," *Nat. Biotechnol.*, vol. 24, pp. 79–88, Jan. 2006.

[33] K. A. Datsenko and B. L. Wanner, "One-step inactivation of chromosomal genes in escherichia coli K-12 using PCR products," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, pp. 6640–6645, June 2000.

[34] R. S. Cox, 3rd, M. J. Dunlop, and M. B. Elowitz, "A synthetic three-color scaffold for monitoring genetic regulation and noise," *J. Biol. Eng.*, vol. 4, p. 10, July 2010.

[35] E. Zorita, P. Cuscó, and G. J. Filion, "Starcode: sequence clustering based on all-pairs search," *Bioinformatics*, vol. 31, pp. 1913–1919, June 2015.

[36] T. Saito and M. Rehmsmeier, "Precrec: fast and accurate precision-recall and ROC curve calculations in R," *Bioinformatics*, vol. 33, pp. 145–147, Jan. 2017.

[37] G. Yu, D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam, "ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data," *Methods Ecol. Evol.*, vol. 8, pp. 28–36, Jan. 2017.

[38] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF chimera–a visualization system for exploratory research and analysis," *J. Comput. Chem.*, vol. 25, pp. 1605–1612, Oct. 2004.

[39] J. A. Capra and M. Singh, "Predicting functionally important residues from sequence conservation," *Bioinformatics*, vol. 23, pp. 1875–1882, Aug. 2007.

[40] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, pp. 2577–2637, Dec. 1983.

[41] J. D. Bloom, "An experimentally determined evolutionary model dramatically improves phylogenetic fit," *Mol. Biol. Evol.*, vol. 31, pp. 1956–1978, Aug. 2014.

[42] T. Izard, "A novel adenylate binding site confers phosphopantetheine adenylyltransferase interactions with coenzyme a," *J. Bacteriol.*, vol. 185, pp. 4074–4080, July 2003.

[43] P. Hu, S. C. Janga, M. Babu, J. J. Díaz-Mejía, G. Butland, W. Yang, O. Pogoutse, X. Guo, S. Phanse, P. Wong, S. Chandran, C. Christopoulos, A. Nazarians-Armavil, N. K. Nasseri, G. Musso, M. Ali, N. Nazemof, V. Eroukova, A. Golshani, A. Paccanaro, J. F. Greenblatt, G. Moreno-Hagelsieb, and A. Emili, "Global functional atlas of escherichia coli encompassing previously uncharacterized proteins," *PLoS Biol.*, vol. 7, p. e96, Apr. 2009.

[44] G. Butland, J. M. Peregrín-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, and A. Emili, "Interaction network containing conserved and essential protein complexes in escherichia coli," *Nature*, vol. 433, pp. 531–537, Feb. 2005.

[45] C. Freiberg, B. Wieland, F. Spaltmann, K. Ehlert, H. Brötz, and H. Labischinski, "Identification of novel essential escherichia coli genes conserved among pathogenic bacteria," *J. Mol. Microbiol. Biotechnol.*, vol. 3, pp. 483–489, July 2001.

[46] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori, "Construction of escherichia coli K-12 in-frame, single-gene knockout mutants: the keio collection," *Mol. Syst. Biol.*, vol. 2, p. 2006.0008, Feb. 2006.

[47] S. Y. Gerdes, M. D. Scholle, M. D'Souza, A. Bernal, M. V. Baev, M. Farrell, O. V. Kurnasov, M. D. Daugherty, F. Mseeh, B. M. Polanuyer, J. W. Campbell, S. Anantha, K. Y. Shatalin, S. A. K. Chowdhury, M. Y. Fonstein, and A. L. Osterman, "From genetic footprinting to antimicrobial drug targets: examples in cofactor biosynthetic pathways," *J. Bacteriol.*, vol. 184, pp. 4555–4572, Aug. 2002.

[48] A. Geerlof, A. Lewendon, and W. V. Shaw, "Purification and characterization of phosphopantetheine adenylyltransferase from escherichia coli," *J. Biol. Chem.*, vol. 274, pp. 27105–27111, Sept. 1999.

Chapter 4

DropSynth 2.0: High-Fidelity Multiplexed Gene Synthesis in Emulsions

4.1 Abstract

Multiplexed assays allow functional testing of large synthetic libraries of genetic elements, but are limited by the designability, length, fidelity and scale of the input DNA. Here we improve DropSynth, a low-cost, multiplexed method which builds gene libraries by compartmentalizing and assembling microarray-derived oligos in vortexed emulsions. By optimizing enzyme choice, adding enzymatic error correction, and increasing scale, we show that DropSynth can build thousands of gene-length fragments at >20% fidelity.

4.2 Main Text

Multiplexed functional assays link gene function or regulation to activities that can be read by next-generation sequencing such as through enrichment screens (cellular growth [1], cell sorting [2, 3], binding [4, 5] or transcriptional reporters [6]. Multiplexed assays can functionally assess thousands of different sequences in a single pooled experiment, and are thus powerful approaches for understanding how sequence affects function [7]. The DNA sequences to test can be accessed

This chapter is an unpublished manuscript that will be submitted for publication as: **A. M. Sidore**[†], C. Plesa[†], N. B. Lubock, D. Zhang, and S. Kosuri "DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions."

by genome fragmentation [8], mutagenesis of existing sequences [9], or by direct synthesis of oligonucleotides (oligos) [10]. Direct oligo synthesis allows for testing controlled hypotheses against one another without the constraints of natural variation or mutagenesis. However, the short length ($<$200nt) of individual oligos that can be synthesized at large scale through array-based approaches limits potential applications. Gene synthesis from oligo libraries can be used to extend these lengths [11, 12], but the high cost of individual assembly and processing becomes prohibitive for large gene libraries.

To address these concerns, we developed a low-cost, multiplexed method termed DropSynth which is capable of building large gene libraries from microarray-derived oligos [13]. DropSynth can build libraries of genes regardless of sequence composition, which enables broad testing of sequence space. However, DropSynth is limited by the resulting fidelity of the gene libraries and the scalability of the method. For example, in our original work, only 1.9-3.9% of assemblies corresponded to the designed protein sequence, and each assembly was limited to 384 designs per library [13].

Here we present DropSynth 2.0, an optimized protocol for multiplexed gene synthesis. We optimize enzyme choice, oligo design, assembly protocols, add enzymatic error correction, and increase the scale that together result in a substantially superior method for gene library synthesis. DropSynth 2.0 works by assembling genes through the isolation and assembly of microarray-derived oligos in droplets (Fig. 4.1A). First, genes are bioinformatically split into several oligos and flanked with restriction sites, priming sequences, and a 12nt microbead barcode sequence that is common to all oligos needed to assemble a given gene (Fig. 4.3). Oligos are synthesized as a microarray-derived pool, amplified and nicked using a nicking endonuclease, exposing each 12nt microbead barcode as a single-stranded overhang. Nicked oligos are hybridized to a pool of barcoded microbeads that contain complementary 12nt microbead barcode sequences, such that each bead pulls down all oligos for a particular assembly. Bound beads are then encapsulated in droplets, where sequences are cleaved from the bead using a IIS restriction enzyme and assembled into genes using a high fidelity polymerase. Following assembly, the emulsion is broken and gene libraries are recovered. Genes possessing mismatches or single-base insertions or deletions contain heteroduplexes, which can be recognized and bound by the bacterial enzyme MutS [14, 15]. Magnetic beads containing immobilized MutS capture these sequences, thus allowing for the enrichment of perfect genes. Finally, gene

libraries are bulk-amplified using single-primer suppression PCR. In this technique, primer annealing competes with the self-annealing of inverted terminal repeats (ITRs) flanking the assembled genes [16, 17]. Shorter by-products tend to self-anneal, while correct assembly products anneal to the primer, resulting in proper amplification.

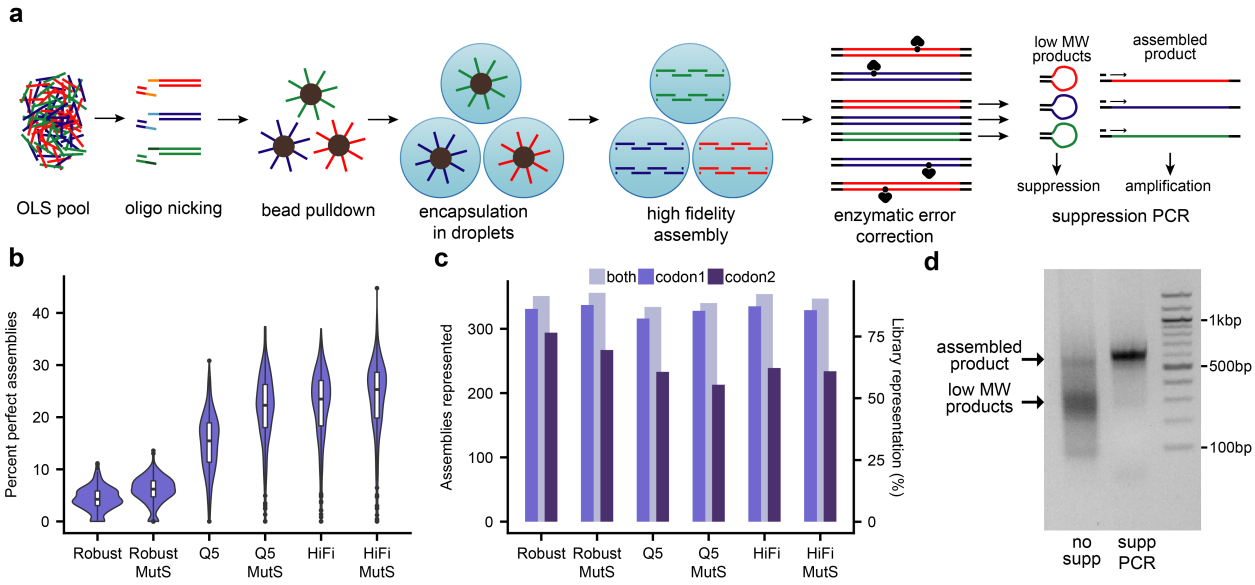

Figure 4.1: **DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions. A.** Schematic of DropSynth 2.0. Refer to Methods for more details. **B.** Comparison of percent perfect assemblies (minimum 100 assembly barcodes) of a 384-gene library assembled using DropSynth with 3 different polymerases (KAPA Robust, NEB Q5, or KAPA HiFi) with or without MutS-based enzymatic error correction. **C.** Comparison of total assemblies represented with at least one assembly barcode for all conditions. 2 codon versions of the 384-gene library were assembled for each condition, and representation is improved when combining across both codon usages. **D.** 2% agarose gel of 384-gene assembly product following bulk amplification with standard PCR or using single-primer suppression PCR; yield of assembled product is noticeably higher using single-primer suppression PCR.

We first set out to validate the efficacy of high-fidelity assembly and error correction in our workflow. To do this, we assembled 2 codon versions of a 384-member library of DHFR homologs using 3 different polymerases (KAPA Robust, NEB Q5, and KAPA HiFi) with or without MutS-based error correction. We ligated the libraries into a plasmid containing a 20bp assembly barcode sequence and sequenced them, allowing us to link assembled genes with unique barcodes. Amongst genes with at least 100 assembly barcodes, we found a median of 4.2% perfect assemblies at the amino acid level for KAPA Robust (Fig. 4.1B), which is consistent with our previous work [13]. We have observed previously that this low fidelity can be attributed to Taq-derived mismatch errors introduced during the assembly [13, 18]. Using high-fidelity polymerases for assembly results in a

several-fold improvement in the median percent perfect assemblies, with 15.5% using NEB Q5 and 23.5% using KAPA HiFi. We also found that using MutS-based error correction resulted in marginal improvements in fidelity (+2.0% for KAPA Robust, +6.8% for NEB Q5, +1.8% for KAPA HiFi) (Fig. 4.1B). A similar trend in percent perfect assemblies was observed from the 2nd codon version assembled (Fig. 4.4).

When analyzing the total number of constructs represented with at least 1 assembly barcode, we found consistently high representation (>75%) across all polymerases for codon 1 (Fig. 4.1C). Interestingly, codon 2 had lower library representation, particularly for NEB Q5 and KAPA HiFi. Though differences in coverage exist between codon usages, combining across codon usages improves the total protein library representation (Fig. 4.1C). Thus, by using multiple codon usages per gene, we improve our ability to achieve greater library coverage. Finally, we observed that using single-primer suppression PCR after assembly significantly improved the quantity of the correctly assembled product, while minimizing the presence of lower molecular weight by-products (Fig. 4.1D).

We next set out to determine algorithmic factors that create differences in library representation. Several factors can contribute to incomplete library representation, including oligo synthesis failure, processing failure, and assembly failure. One cause of assembly failure is the inability of oligos to overlap and assemble properly. In order to investigate this further, we created multiple iterations of the same 2 codon versions of our 384-member DHFR library using different overlap parameters, including overlap length and secondary structure [19]. We found that 20bp overlaps had higher library representation than 25bp overlaps, while modifying the secondary structure had minimal effect (Fig. 4.5).

Assembly failure can also be attributed to incompatibilities between the polymerase buffer and the IIS restriction enzyme used to cleave oligos off the beads. In particular, NEB Q5 buffer inhibits several IIS restriction enzymes [20], which can cause incomplete library representation by preventing the cleavage of oligos from the surface of the microbead within the droplet (Fig. 1c). To investigate this further, we designed multiple iterations of the same 2 codon usages of our 384-member DHFR library with 3 different IIS restriction sites (BtsI, BsmAI and BsrDI) and assembled them using NEB Q5. Though differences in library representation exist across codon versions, we found that assemblies using BsrDI had poor representation when compared to assemblies with BtsI and BsmAI.

(Fig. 4.1).

The scale at which we can build gene libraries using DropSynth is currently limited to 384 genes per reaction. In an effort to overcome this limitation, we designed and created a new barcoded bead pool containing 1536 unique microbead barcode sequences. This new bead pool was constructed using similar procedures to the 384-plex bead pool (Fig. 4.1). In order to demonstrate the efficacy of the new bead pool, we designed and assembled 2 codon versions of a 1536-member library of DHFR homologs. Each library member contains one of 1536 unique microbead barcode sequences which can be hybridized to one of 1536 beads with complementary barcode sequences. We assembled these libraries using KAPA HiFi and ligated them into a barcoded expression plasmid. Following sequencing, we observed 1048/1536 (codon 1) and 904/1536 (codon 2) constructs represented with at least one assembly barcode (Fig. 4.2A). When combining across codon usages, we found a total of 1208 constructs represented, approaching 80% total protein library coverage (Fig. 4.2A). Amongst genes with at least 100 assembly barcodes, we found a median of 27.6% perfect assemblies for codon 1 and 22.6% for codon 2, suggesting that the new bead pool can assemble large libraries at high fidelity (Fig. 4.2B).

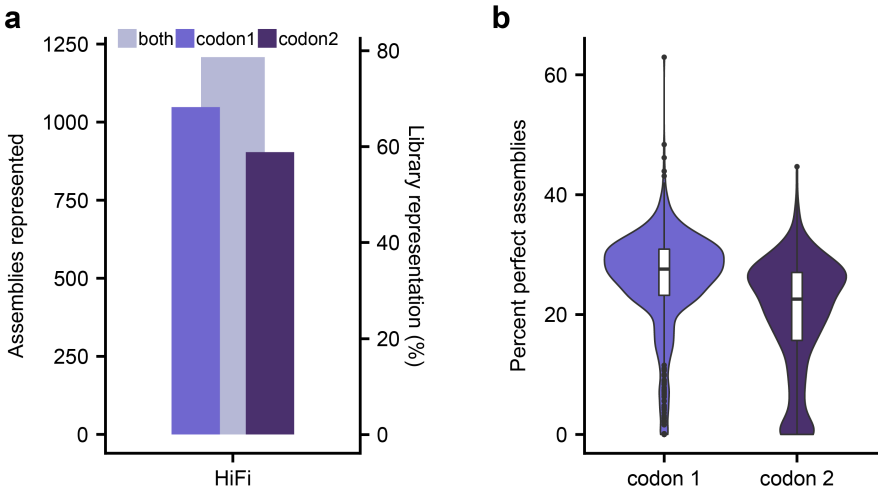

Figure 4.2: **A scaled-up barcoded bead pool allows for the one-pot assembly of up to 1536 genes. A.** 2 codon versions of a 1536-gene library were assembled using KAPA HiFi; when combining across both codon usages, 1208/1536 genes have at least one assembly barcode. **B.**Comparison of percent perfect assemblies (minimum 100 assembly barcodes) of both codon versions of each 1536-gene library.

DropSynth 2.0 combines improvements in fidelity and scale, significantly enhancing our ability to build large, accurate gene libraries. By improving fidelity, gene libraries enriched with perfect

assemblies enable clearer hypothesis testing using multiplexed functional assays. In addition, improvements in fidelity allow for the assembly of longer genes using more oligos. Finally, increasing the frequency of perfect assemblies enables simpler individual gene retrieval using molecular cloning or dial-out PCR [21]. By improving scale, larger gene libraries reduce the cost per assembly reaction and enable more data to be generated on desired hypotheses. Combining these improvements creates a much more powerful workflow for the synthesis of large gene libraries.

4.3 Materials and Methods

Oligo design

The software used to split a given amino acid sequence into oligos with overlaps was derived from Eroshenko et al19 and available on https://github.com/KosuriLab/. Amino acid sequences were first assigned random weighted codons based on their frequency in the E. coli genome and screened against several illegal restriction sites (NdeI, KpnI, BtsI-v2, and BspQI). Next, the coding regions were flanked with restriction sites for cloning (NdeI, KpnI) and 20mer assembly primers used in the emulsion polymerase cycling assembly (PCA). The sequences were then split into oligos with overlap regions obeying certain parameters, including melting temperature range, mean overlap size and secondary structure. Sequences that failed to meet these parameters were assigned new codons until a successful split was generated. Split oligo sequences were then flanked with BtsI-v2 sites used to release the oligos inside each droplet. In order to maintain the same length across all oligos, padding sequence consisting of ATGC repeats was added to the region upstream of the 5' BtsI-v2 site. Next, a Nt.BspQI sequence, 12mer gene-specific barcode sequence (referred to as the 'microbead barcode'), and another Nt.BspQI sequence were prepended to the 5' end of each oligo. Nt.BspQI was used to nick the top strand on the 5' end of the barcode and the bottom strand on the 3' end of the barcode sequence, exposing it as a 12nt top-strand overhang. This barcode allows all oligos contributing to a given gene to be localized on the same bead. Oligos were next flanked with 15mer amplification primers unique to a given library subpool. BLAT [22] was run to verify that amplification primer sequences did not possess homologies >10bp to designed oligos. Prior to synthesis, final oligo sequences were screened for the presence of all required components and against

all illegal restriction sites.

Using the above oligo design, we synthesized a microarray-derived oligo library synthesis (OLS) pools of 33,792 230mer oligos from Agilent Technologies. This pool contained several variations of two codon versions of a 384-member DHFR library derived from our original work13. For our control libraries, which were used for all biological optimizations, we used an overlap melting temperature range of 58°C-62°C, mean overlap size of 20bp, and an overlap secondary structure cutoff of -4 kcal/mol. We also generated identical amino acid libraries using alternative overlap parameters, including a longer overlap size of 25bp, and a more stringent secondary structure cutoff of -2 kcal/mol. Another set of amino acid libraries contained alternative IIs restriction sites to BtsI-v2, including BsmAI and BsrDI. This OLS pool also contained 2 codon usages of a single 1536-member DHFR library derived from 4 libraries from Plesa et al [13].

Microbead barcode design

In order to generate distinct 12mer barcode sequences, we took 2,000 20mer primer sequences derived from Eroshenko et al [19], removed all sequences containing NdeI, KpnI, BtsI-v2, BspQI, EcoRI, XhoI, SpeI, and NotI, and generated all possible 12mer subset sequences. We next screened for self-dimers, GC content between 45% and 55% and a melting temperature between 40°C and 42°C, We further filtered sequences to have a minimum modified Levenshtein distance of 3 between selected barcodes23. We then selected the first 384 sequences to be used in oligo designs, with complementary sequences being used to generate the beads. For the 1,536-plex barcode design, we performed identical screens except for a relaxed melting temperature screen between 38°C and 44°C. The first 1,536 sequences were used in our 1,536-plex oligo libraries, with complementary sequences being used to generate the beads.

Barcoded beads protocol

Three oligos are required to generate each DropSynth barcoded bead, two of which are common to all beads (anchor and ligation oligo). The anchor oligo, which has 5' double biotin modification, contains sequences complementary to the ligation oligo and part of the barcode oligo. The ligation oligo, which contains 3' biotin modification and 5' phosphate modification, is fully complementary

to the anchor oligo and allows for the ligation of the barcode oligo. The barcode oligo, which has no modifications, contains a common sequence on the 3' end which hybridizes to the anchor oligo, and a unique 12nt sequence which acts as a 5' overhang. This setup minimizes cost, as only the common oligos (anchor and ligation) require expensive modifications. The anchor and ligation oligo were purchased in bulk at >1umole while the barcode oligos were purchased as a single 384-well plate from Integrated DNA Technologies.

The three oligos required for each barcoded bead were individually mixed, ligated, and phosphorylated in individual wells of a 384-well plate using a Liquidator 96 (Rainin). Next, magnetic Streptavidin M270 Dynabeads (Invitrogen) were added to each well, and plates were incubated overnight at room temperature while shaking >2000RPM. The individual wells were then washed textgreater5 times using 2X Bind & Wash Buffer and a 384-Well Post Magnetic Plate (Permagen). After washing, individual bound beads were resuspended in 5ul of Bind & Wash Buffer and pooled together. For the 1536-plex barcoded bead pool, 4 plate pools of 384 barcoded beads were combined in equal volumes.

Oligo amplification and processing

Upon receipt of the oligo pool, individual oligo libraries were PCR-amplified using 15mer amplification primers with Q5 High-Fidelity 2X Master Mix (New England Biolabs), and number of cycles determined by qPCR. Amplifications were stopped several cycles prior to plateauing to prevent overamplification. Oligo subpools were then diluted to 0.02ng/ul and bulk-amplified using a biotinylated forward amplification primer and unmodified reverse amplification primer with Q5 High-Fidelity 2X Master Mix for 20 cycles. For each library, 8 PCRs were run in parallel, pooled and column-cleaned using a Zymo Clean & Concentrator. Oligo subpools were then nicked overnight using the nicking endonuclease Nt.BspQI, exposing gene-specific 12nt barcode overhangs. The short biotinylated fragment cleaved following nicking was removed by binding to Streptavidin M270 Dynabeads (Invitrogen), and the remaining processed oligos were column-cleaned. 1.3 ug of each processed oligo subpool was added to 20ul of barcoded beads (5 million beads) and Taq ligase. The mixture was slowly annealed overnight from 50°C to 10°C, allowing the 12nt overhang on the processed oligos to hybridize to complementary 12nt overhangs on barcoded beads.

Emulsion assembly

Loaded beads were mixed with a polymerase master mix (KAPA 2G Robust HotStart ReadyMix, KAPA HiFi HotStart ReadyMix, or Q5 High-Fidelity 2X Master Mix), 60mer primer sequences containing 20nt amplification primer sequences and 40nt ITRs (to be used during bulk suppression PCR), BSA, and BtsI-v2. Immediately after adding BtsI-v2, the mixture was added to 600ul of BioRad Droplet Generation Oil and vortexed for 3 minutes using a Vortex Genie 2 (Scientific Industries), resulting in compartmentalization of beads in <5um droplets. After vortexing, samples were aliquoted into PCR strips and incubated at 55C for 90 minutes, allowing BtsI-v2 to cleave oligo sequences off the beads. Samples were then thermocycled for 60 cycles, allowing polymerase cycling assembly to proceed. Emulsions were broken by adding 100ul perfluoro-1-octanol, and the aqueous phase was extracted and column-cleaned. Assembled products were then run on a 2% agarose gel and bands were extracted at the correct assembly length.

Mismatch binding by MutS

Following gel extraction of assembly products, 10ul of M2B2 magnetic beads (US Biological) was added to each library and incubated for 2 hours at room temperature while shaking using a Thermomixer C (Eppendorf). M2B2 beads contain immobilized MutS and thus bind to and magnetically separate DNA containing mismatch-generated heteroduplexes. Following incubation, error-depleted libraries were column-cleaned using a Zymo Clean & Concentrator. In order to verify filtration of DNA, libraries were bulk-amplified on a qPCR using assembly primers before and after M2B2 treatment and deltaCq was quantified.

Bulk suppression PCR

Gene libraries assembled during DropSynth assembly contain external 40bp inverted terminal repeats (ITRs) lacking homology to any library sequences. Following recovery of assembled DropSynth libraries, a bulk PCR was carried out using a single 20nt primer complementary to the proximal region of the 5' ITR. Due to their close physical proximity, the ITRs of shorter DNA fragments tend to self-anneal, creating hairpin-like structures with suppressed amplification. In contrast, the ITRs of longer DNA fragments are less likely to anneal to one another, allowing for primer annealing and

effective amplification. In this case, libraries were amplified using Q5 High-Fidelity 2X Master Mix (New England Biolabs), a final primer concentration of 0.8uM, Tm of 58°, and number of cycles determined by qPCR. Amplifications were stopped several cycles prior to plateauing to prevent overamplification. Following amplification, samples were run on a 2% agarose gel and assembly bands were extracted.

pEVBC plasmid construction

The plasmid used to barcode unique assemblies is derived from our previous work [13]. pEVBC is a pUC19 derivative containing a pLac-UV5 promoter, NdeI and KpnI restriction sites for cloning, an in-frame stop codon and 20mer random assembly barcode sequences. The plasmid was constructed by digesting pUC19 with AatII and BspQI, gel-extracting the larger fragment, and ligating in a gBlock DNA fragment containing the promoter, several restriction sites, and chloramphenicol acetyltransferase in frame before the stop codon. The resulting plasmid was then double digested with NcoI and KpnI and the 2,209bp fragment was gel extracted. Using this fragment as a template, an around-the-horn PCR was carried out using the forward primer pEVBC-FWD containing an NdeI site and reverse primer pEVBC-REV1 containing KpnI and a 20mer random assembly barcode sequence for 5 cycles. The PCR product was then further amplified using pEVBC-FWD and pEVBC-amp-FWD for 15 cycles. The resulting amplicon was then column purified, digested with NdeI and KpnI, treated with rSAP and size-selected.

Barcoded library in pEVBC

Following bulk suppression PCR of assembly products, gene libraries were double-digested with NdeI and KpnI and column-purified. Gene libraries were then ligated to digested NdeI + KpnI pEVBC plasmid using a 3:1 insert-to-vector molar ratio, column-purified, and eluted in a volume of 15ul. Ligation products were directly PCR-amplified with sequencing primers mi3-FWD and mi3-N#-REV to add p5, p7, and indexes for Illumina sequencing.

Assembly barcode sequencing and analysis

Assembly barcoded libraries were sequenced on a total of 5 Illumina MiSeq paired-end 600-cycle runs. Following PCR amplification with sequencing primers mi3-FWD and mi3-N7#-REV, amplicons were gel-extracted and quantified using an Agilent 2200 TapeStation. Samples were then pooled and sequenced on a MiSeq using custom primers mi3-R1, mi3-R2 and mi3-ndex, and fastqs were generated for each sample following demultiplexing. In order to eliminate biases in coverage following sequencing, individual fastqs were randomly downsampled to 1,880,288 reads (number of reads of the sample with the lowest read depth). All fastq files were trimmed of adapter sequences with bbduk, and paired-end reads were merged with bbmerge (from BBTools package). Reads were next concatenated and piped into a custom python script, used in our previous work. This script splits reads into variants and 20nt assembly barcodes, generating a dictionary containing each assembly barcode and the variants mapped to it. Assembly barcodes that map to multiple variants were removed by calculating the pairwise Levenshtein distance of every variant associated with a given assembly barcode. If at least 5% of assembly barcodes have a Levenshtein distance >10, the assembly barcode is considered contaminated and dropped from the analysis. Next, a consensus sequence is generated by taking the majority base call at each position, and translated until the first stop codon. Variants and their mapped barcodes were then imported into R, where they were analyzed for coverage and fidelity. For coverage analyses, the term 'assemblies represented' refers to the total number of assemblies corresponding to a perfect amino acid sequence represented by at least one assembly barcode. For fidelity analyses, the term 'percent perfect assemblies' is defined as the median percent perfect sequences at the amino acid level determined by using constructs with at least 100 assembly barcodes.

DropSynth 2.0 bead barcoding protocol

This protocol can be performed using 1 384-well plate to generate 384 unique barcoded beads, or 4 384-well plates to generate 1536 unique barcoded beads. Though the process can be done by hand, it is helpful to use a Rainin Liquidator 96 for liquid handling steps.

Reagents Required (384-plex):

- 240 uL 100 uM anchor oligo (Integrated DNA Technologies)
- 240 uL 100 uM ligation oligo (Integrated DNA Technologies)
- 1,056 uL 10X T4 Ligase Buffer (New England Biolabs)
- 1 uL of each 100 uM barcoded oligo (Integrated DNA Technologies)
- 24 uL T4 Ligase (New England Biolabs)
- 240 uL T4 PNK (New England Biolabs)
- 1500 uL Streptavidin M270 Dynabeads (Invitrogen)
- >10 mL UltraPure Distilled Water (Invitrogen)
- >10 mL 2X B&W Buffer

Reagents Required (1536-plex):

- 960 uL 100 uM anchor oligo (Integrated DNA Technologies)
- 960 uL 100 uM ligation oligo (Integrated DNA Technologies)
- 4,224 uL 10X T4 Ligase Buffer (New England Biolabs)
- 1 uL of each 100 uM barcoded oligo (Integrated DNA Technologies)
- 96 uL T4 Ligase (New England Biolabs)
- 960 uL T4 PNK (New England Biolabs)
- 6,000 uL Streptavidin M270 Dynabeads (Invitrogen)
- >40 mL UltraPure Distilled Water (Invitrogen)
- >40 mL 2X B&W Buffer

Prepare 40mL 2X B&W buffer (2M NaCl, 1mM EDTA, 10mM Tris):

- 4.675g NaCl salt
- 400 uL UltraPure 1M Tris, pH 7.5 (Invitrogen)
- 80 uL UltraPure 0.5 M EDTA, pH 8.0 (Invitrogen)
- UltraPure Distilled Water (Invitrogen) to 40 mL

1. Hybridize the anchor, ligation and barcoded oligos:
    - Add to the first row of 96-well deep well plate:
        - 20 uL 100 uM anchor oligo
        - 20 uL 100 uM ligation oligo
        - 80 uL 10X T4 Ligase Buffer
        - 640 uL UltraPure Distilled Water
    - Using a Rainin P200 12-channel pipette, add 95 uL of master mix to all rows of master 96-well plate.
    - Using a Rainin Liquidator 96, distribute 19 uL of master mix from master 96-well plate to all wells of a new 384-well plate. The 384-well plate can be adjusted to 4 corners using a Rainin Plate Adapter 384, allowing all wells to be filled from the 96-well master plate.
    - Using a Rainin Liquidator 96, transfer 1 uL from every well of the 100 uM barcoded oligo plate to every well of the 384-well plate.
    - Anneal the mixed oligos on each plate using the following conditions:
        - 3 min at 70°C
        - Ramp down to 60°C for 1 min, 0.1°C/sec
        - Ramp down to 50°C for 1 min, 0.1°C/sec
        - Ramp down to 40°C for 1 min, 0.1°C/sec
        - Ramp down to 30°C for 1 min, 0.1°C/sec
        - Put plate on ice

2. Ligate the barcoded oligo to the ligation oligo:

    - Add to the first row of a 96-well plate:
        - 2 uL T4 Ligase
        - 8 uL 10X T4 Ligase Buffer
        - 70 uL UltraPure Distilled Water
    - Using a Rainin P20 12-channel pipette, add 10 uL of master mix to all rows of a master 96-well plate.
    - Using a Rainin Liquidator 96, distribute 2 uL master mix from the master 96-well plate to all wells of the 384-well plate.
    - Incubate plate at 16°C for 1 hr or longer, followed by 65°C for 20 min to heat inactivate the ligase

3. Phosphorylate the barcoded oligo:

    - Add to first row of 96-well plate:
        - 20 uL T4 PNK
        - 60 uL UltraPure Distilled Water
    - Using a Rainin P20 12-channel pipette, add 10 uL of master mix to all rows of a master 96-well plate.
    - Using a Rainin Liquidator 96, distribute 2 uL master mix from the master 96-well plate to all wells of the 384-well plate.
    - Incubate the plate at 37°C for 40 min (or longer), followed by 65°C for 20 min to heat inactivate the PNK

4. Bind to beads:

    - Prepare 1500 uL stock Dynabeads M270 Streptavidin, washed, and resuspended in 3000 uL 2X B&W buffer.
    - Add 200 uL to first row of 96-well plate.

- Using a Rainin P200 12-channel pipette, add 25 uL of master mix to all rows of a master 96-well plate.
- Using a Rainin Liquidator 96, add 5 uL resuspended beads to each well of the 384-well plate.
- Mix overnight with shaking (>2000 RPM) at room temperature.

5. Pool beads:

   - Using a Rainin Liquidator 96, wash each well with 20 uL 2X B&W buffer 8 times.
   - Using a Rainin Liquidator 96, resuspend each well in 5 uL 2X B&W buffer.
   - Mix 5 uL of each well together, making a 1920 uL mixed barcoded bead pool for each plate. Store these at 4°C when not in use.

DropSynth 2.0 emulsion synthesis protocol

1. Prepare the OLS pool

   - Make a 1/10 dilution of the OLS chip pool.
   - Prepare mixtures of forward and reverse subpool amplification primers for each subpool, with 10 uM final concentration of each primer.

2. Amplify subpools.

   - For each subpool, run a qPCR to determine the number of cycles required for amplification. Amplifications are stopped several cycles before plateauing to prevent over-amplification of the libraries.
   - Amplify each subpool using NEB Q5.
     - 1 uL template (1/10 OLS pool dilution)
     - 1.25 uL subpool specific primer 10 uM ampF
     - 1.25 uL subpool specific primer 10 uM ampR
     - 21.5 uL UltraPure Distilled Water (Invitrogen)

- 25 uL NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs)
- TOTAL: 50 uL
- PCR protocol:
  1. 45 sec 98°C initial denaturation
  2. 15 sec 98°C denaturation
  3. 30 sec 58°C annealing
  4. 15 sec 72°C extension
  5. Go to step 2, repeat based on the number of cycles determined by qPCR.
  6. 1 min 72°C final extension

- Column purify amplified oligos using a DNA Clean & Concentrator -5 (Zymo Research).
- Run PCR products on gel. Look for higher MW products, indicative of overamplification. Excessive low MW products may indicate chip synthesis issues.
- Size select, using gel extraction, if necessary.
- Create 20 pg/uL dilutions of each amplified subpool.

3. Bulk amplify subpools.

- Run a second PCR using a biotinylated FWD amplification primer, with sufficient tubes to make 5 ug to 10 ug of PCR product.
  - 1 uL of 20 pg/uL subpool dilution
  - 1.25 uL subpool specific primer mix 10 uM biotinylated ampF
  - 1.25 uL subpool specific primer mix 10 uM biotinylated ampR
  - 21.5 uL UltraPure Distilled Water (Invitrogen)
  - 25 uL NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs)
  - TOTAL: 50 uL
  - PCR protocol:
    1. 45 sec 98°C initial denaturation
    2. 15 sec 98°C denaturation

3. 30 sec 58°C annealing
4. 15 sec 72°C extension
5. Go to step 2, 18X
6. 1 min 72°C final extension

- Pool and column purify using a DNA Clean & Concentrator -25 (Zymo Research).

4. Nicking.

- Nick the bulk amplified subpools. Split the following across multiple tubes depending on the amount of DNA to be processed. In each 1.5 mL tube add:
  - 15 uL Nt.BspQI (10U/uL) (New England Biolabs)
  - 5 to 10 ug of DNA
  - 15 uL NEBuffer3.1 (New England Biolabs)
  - UltraPure Distilled Water (Invitrogen) to 150 uL total
- Leave at 50°C overnight with shaking >1500 RPM.

5. Capture and remove the short biotinylated fragment.

- Wash 50 uL streptavidin M270 Dynabeads (Invitrogen) for each 1.5 mL tube in the nicking reaction, as per manufacturer's instructions and resuspend in 2X B&W buffer.
- Add 50 uL of washed beads to the 150 uL nicking reaction in each tube.
- Incubate at 55°C with 800 RPM shaking for at least 1 hour.
- Move all 1.5 mL tubes to a 55°C water bath.
- Place the tube so that solution is just below the surface of the water. Hold a strong magnet underwater against the side of the tube to magnetically separate Dynabeads. Pipette the supernatant, which contains the processed oligos and save them in a new container. Remove the tube with the Dynabeads from the magnet.
- Add 100 uL of UltraPure Distilled Water (Invitrogen) to the tube and resuspend the beads. Incubate these at 55°C for another 30 min and then repeat the procedure to recover the supernatant again while leaving the Dynabeads behind.

- Repeat this procedure for all tubes as necessary.
- Pool processed oligos (supernatant) for each subpool and column purify using a DNA Clean & Concentrator -5 (Zymo Research).

6. Capture processed oligos with barcoded beads.
   - Take 20 uL of the pooled barcoded beads. These are in stored in 2X B&W buffer (high ionic concentration) which may interfere with ligation reaction. Resuspend them in 20 uL UltraPure Distilled Water (Invitrogen).
   - Mix the processed DNA with the barcoded beads:
     - 1.3 ug processed DNA (12 pmol)
     - 20 uL pooled barcoded beads (6 million beads, binding capacity 1.3 ug DNA)
     - 10 uL 10X Taq ligase buffer (New England Biolabs)
     - 4 uL Taq ligase (40 U/uL) (New England Biolabs)
     - UltraPure Distilled Water (Invitrogen) to 100 uL
   - Overnight cycling (>2 hr incubation at each of the following temperatures) (13 hr), use shaking to prevent beads from settling down:
     - 3 hours @ 50°C
     - Ramp to 40°C for 3h, 0.1°C/sec
     - Ramp to 30°C for 3h, 0.1°C/sec
     - Ramp to 20°C for 2h, 0.1°C/sec
     - Ramp to 10°C for 2h, 0.1°C/sec
   - Wash 3 times at 4°C using 2X B&W buffer. This is important for removing unbound oligos in order to increase specificity.
   - Wash twice at RT using 2X B&W buffer
   - Re-suspend in 100 uL Elution Buffer (Qiagen) (60k beads/uL)

7. Emulsion assembly (ePCA).

- Setup emulsion. All of this procedure should be done on ice. FWD and REV assembly primers contain ITR overhangs which will be used for single-primer suppression PCR. Add BtsI-v2 only at the very last step. Try to minimize the time between adding the BtsI-v2 and vortexing the emulsion.
  - 40 uL of loaded beads (500 ng DNA)
  - 0.5 uL 100 uM AsmF-40bpITR
  - 0.5 uL 100 uM AsmR-40bpITR
  - 50 uL KAPA HiFi 2X Mastermix (KAPA Biosystems)
  - 1 uL BSA (New England Biolabs)
  - 1 uL UltraPure Distilled Water (Invitrogen)
  - 7 uL BtsI-v2 (New England Biolabs) (add last)
  - TOTAL: 100 uL
- Mix at low speed in vortexer to resuspend beads.
- Add 600 uL Droplet Generation Oil for EvaGreen (Bio-Rad) to a 1.5mL non-stick tube.
- Add 100 uL aqueous phase to the bottom of the oil phase.
- Vortex at Max Speed in foam holder taped down for 3 minutes. If doing multiple emulsions, do this one at a time. We use a Vortex Genie 2 (Scientific Industries) at max speed.
- After vortexing all emulsions, place each emulsion into PCR tubes with 100 uL in each tube. Use a P1000 tip to avoid disturbing the emulsion. Most of the droplets will float to the top of the tube, try to get as much of this as possible and distribute this over multiple PCR tubes.
- PCR Cycling
  1. 55°C for 90 min (allow BtsI-v2 to cleave DNA from the beads)
  2. 94°C for 2 min (initial denaturing)
  3. 94°C for 15 sec (denaturing)
  4. 57°C for 20 sec (annealing)
  5. 72°C for 45 sec (extension)

6. Go to step 3 for additional 60 cycles
7. 72°C for 5min (final extension)
8. 4°C forever

8. Break the emulsion:

- Pipet out the entire volume of droplets from each PCR tube into a 1.5 mL tube.
- Add 100 uL of 1H,1H,2H,2H-Perfluoro-1-octanol (Sigma Aldrich) for each 100 uL of PCR reaction combined in the 1.5mL tube.
- Vortex at maximum speed for 1 min.
- In a centrifuge, spin down at 15,500 x g for 10 min.
- If droplets are still present, vortex and centrifuge again.
- Remove upper aqueous phase by pipetting, avoiding the oil phase.
- Transfer this to a clean 1.5mL tube (this is the DNA).
- Column purify using a DNA Clean & Concentrator -5 (Zymo Research).

9. Size selection via gel extraction.

- Run amplicons on a gel and extract the correct range and purify.
- Note: Typically there is not enough DNA after the ePCA to visualize on a gel, so this is often a blind extraction.

10. MutS treatment (optional)

- Enzymatic error correction can be used to enrich for perfect assembly products. Here we use M2B2 magnetic beads (US Biological), which contain immobilized MutS and thus bind to and magnetically separate DNA containing mismatch-generated heteroduplexes.
- Add 10 uL of M2B2 magnetic beads to size-selected assembly product.
- Incubate at 20°C with 1600 RPM shaking for at least 1 hour.
- Immediately place on magnetic rack and extract supernatant.
- Column clean the DNA using a DNA Clean & Concentrator -5 (Zymo Research).

11. Single-primer suppression PCR.

    - In this technique, self-annealing of inverted terminal repeats (ITRs) flanking the assembled genes competes with the annealing of a single primer which aligns to part of the ITR3. Shorter by-products tend to self-anneal, while correct assembly products anneal to the primer, resulting in proper amplification.
        - 1 uL template
        - 4 uL 10 uM suppression primer
        - 25 uL NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs)
        - UltraPure Distilled Water (Invitrogen) to 50 uL
        - PCR protocol:
            1. 45 sec 98°C initial denaturation
            2. 15 sec 98°C denaturation
            3. 30 sec 58°C annealing
            4. 15 sec 72°C extension
            5. Go to step 2, determine cycles using qPCR.
            6. 1 min 72°C final extension
    - Column purify using a DNA Clean & Concentrator -5 (Zymo Research).
    - Check size distribution on gel or tapestation.
    - Quantify DNA and proceed to downstream applications.

4.4 Supplementary Information

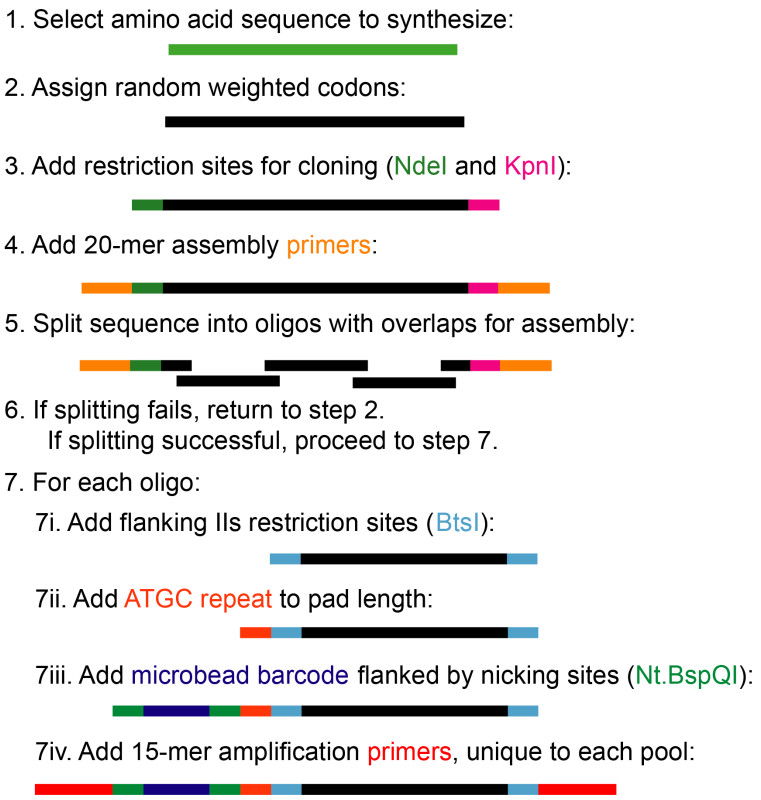

Figure 4.3: **Overview of the DropSynth oligo design process.** The oligo design script, available at https://github.com/KosuriLab/DropSynth and originally derived from Eroshenko et al. [19], takes as input a list of protein sequences and generates all oligos necessary to assemble each gene. First, amino acid sequences are assigned random weighted codons and flanked with restriction sites used for cloning and 20mer assembly primer sequences used for the emulsion assembly. Next, the full gene sequence with restriction sites and primers is split into oligos with overlaps of a predefined length, melting temperature and secondary structure. If splitting fails, which can be due to improper overlap parameters, long homopolymers, or illegal restriction sites, the protein sequence is reassigned new random weighted codons and the process is repeated. Once each gene is successfully split into oligos, each oligo is flanked with BtsI sites used to cleave sequences off beads, padding sequence, a 12mer gene-specific microbead barcode sequence flanked by Nt.BspQI sites, and 15mer amplification primer sequences used to amplify the oligo libraries from the OLS pool.

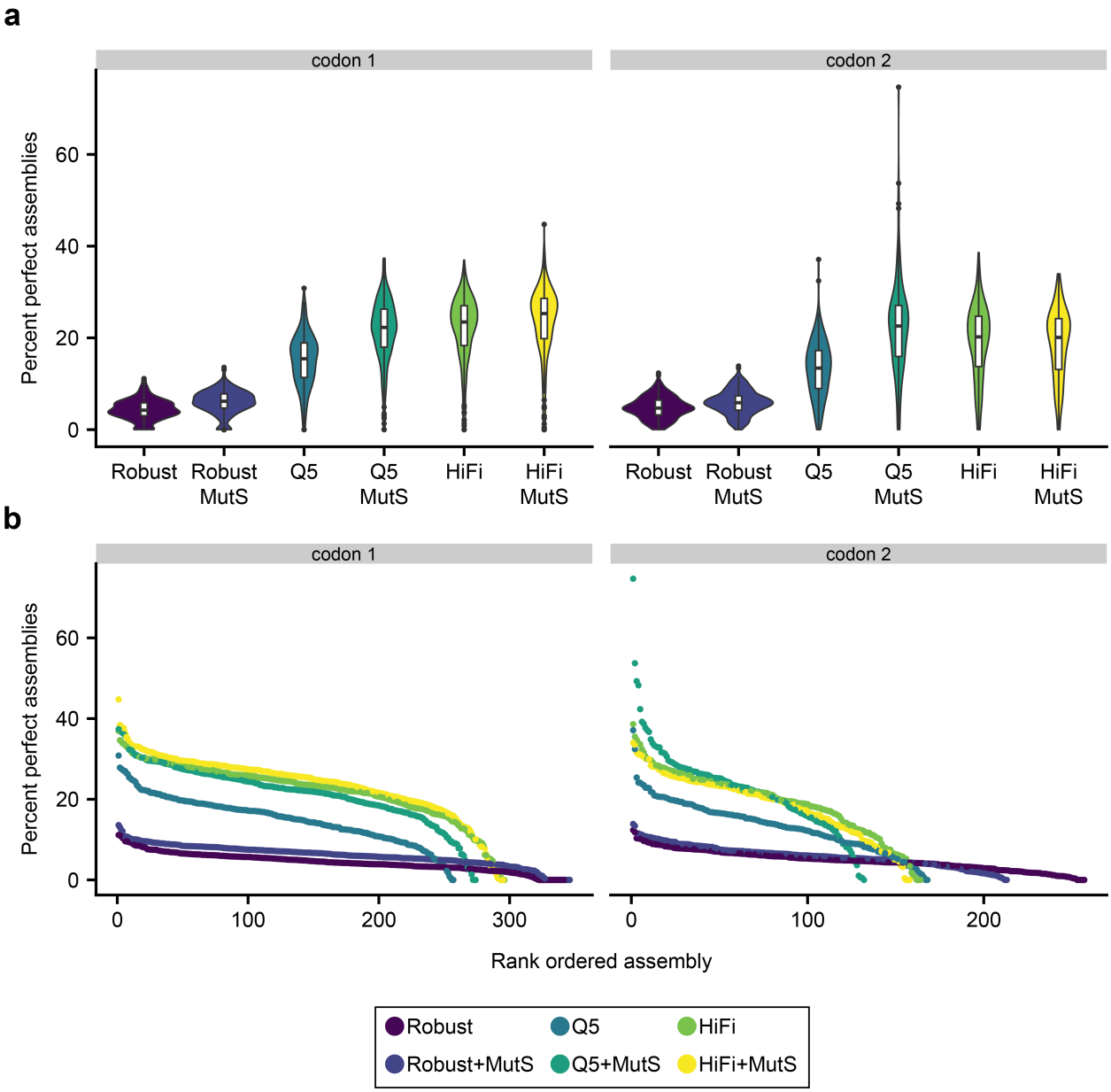

Figure 4.4: **DropSynth assembly of 2 codon versions of a 384-gene library using 3 different polymerases with or without MutS-based enzymatic error correction. A.** Comparison of percent perfect assemblies (minimum 100 assembly barcodes) of 2 codon versions of a 384-gene library assembled using DropSynth with 3 different polymerases (KAPA Robust, NEB Q5, or KAPA HiFi) with or without MutS-based enzymatic error correction. **B.** Rank ordered plot of percent perfect assemblies (minimum 100 assembly barcodes) of all conditions. Though assemblies with KAPA Robust have the greatest library representation, assemblies with high-fidelity polymerases NEB Q5 and KAPA HiFi have significantly improved fidelity of represented constructs.

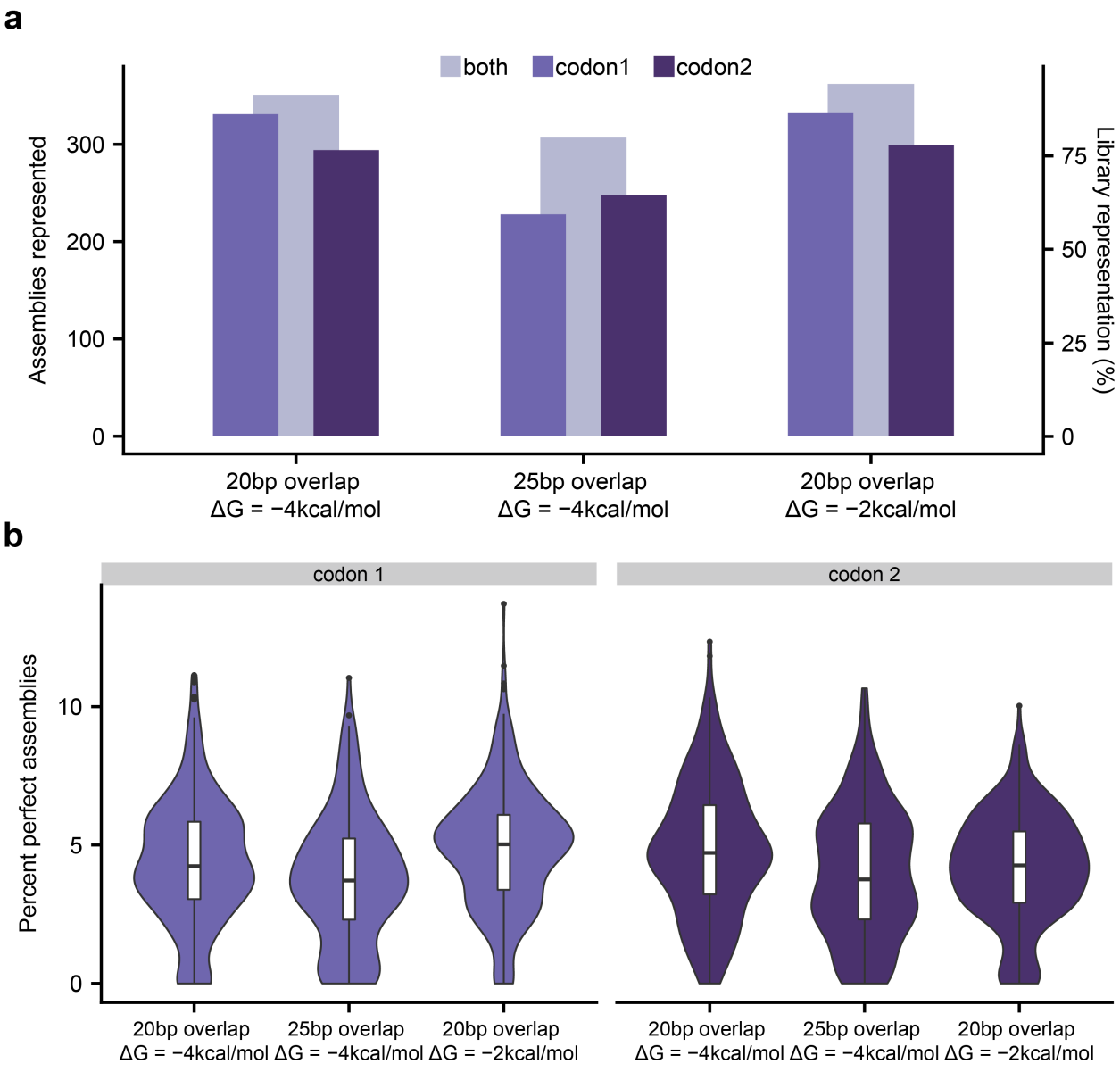

Figure 4.5: **DropSynth assembly of 2 codon versions of a 384-gene library containing alternative oligo overlap parameters (length, secondary structure). A.** Comparison of total assemblies represented with at least one assembly barcode of 2 codon versions of a 384-gene library designed with alternative average overlap lengths (20 or 25bp) and overlap secondary structure thresholds (maximum deltaG = -4 kcal/mol or -2 kcal/mol) and assembled using DropSynth with KAPA Robust. Modifying the overlap secondary structure appears to have little effect on representation, while increasing the average overlap length to 25bp has a slight negative effect on representation. **B.** Comparison of percent perfect assemblies (minimum 100 assembly barcodes) of all conditions.

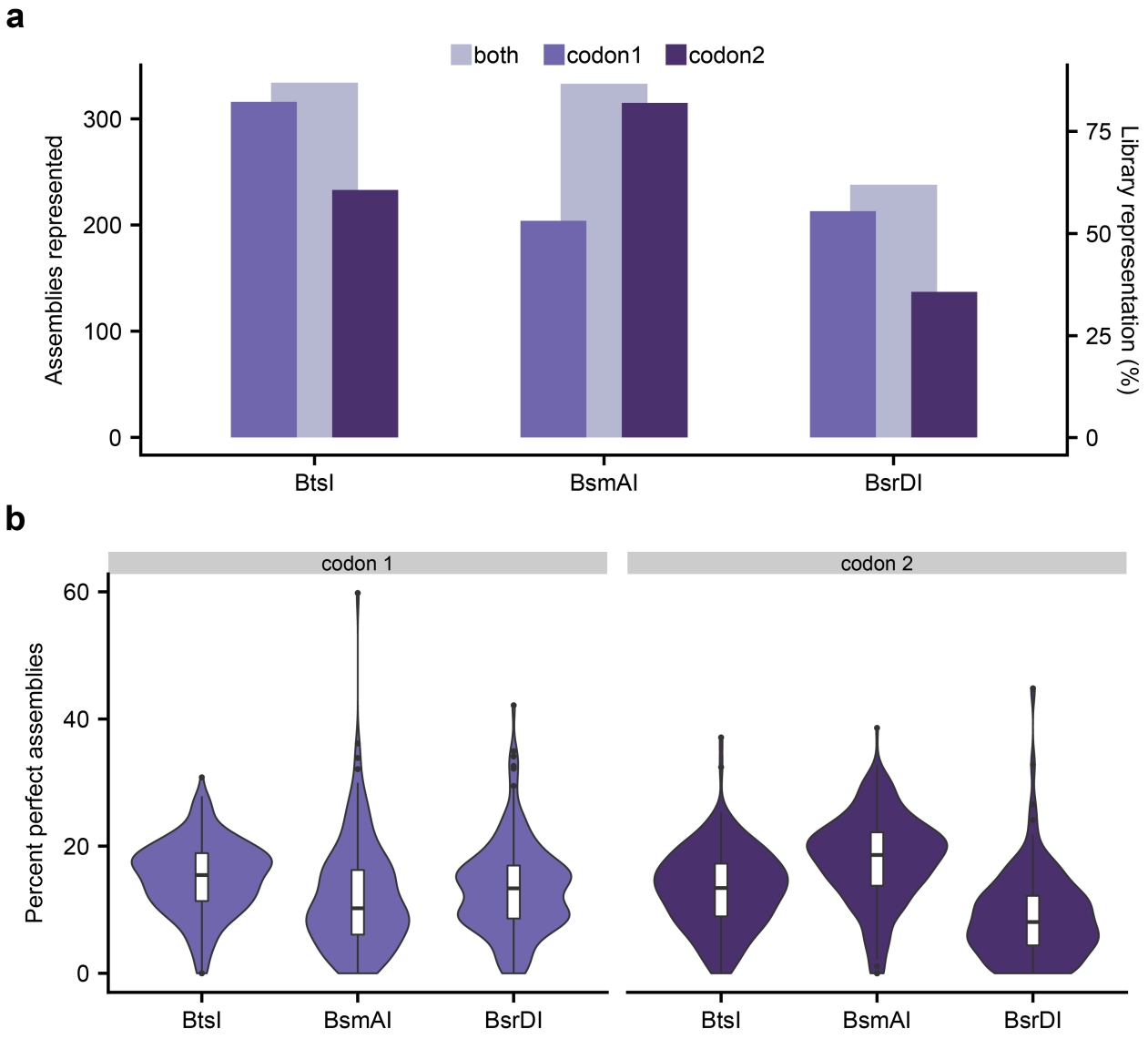

Figure 4.6: **DropSynth assembly of 2 codon versions of a 384-gene library containing alternative IIS restriction sites (BtsI, BsmAI, and BsrDI). A.** Comparison of total assemblies represented with at least one assembly barcode of 2 codon versions of a 384-gene library designed with alternative IIS restriction sites used to cleave oligos off the beads (BtsI, BsmAI, or BsrDI) and assembled using DropSynth with NEB Q5. Using BsrDI appears to have a slight negative effect on representation compared to BtsI and BsmAI. **B.** Comparison of percent perfect assemblies (minimum 100 assembly barcodes) of all conditions.

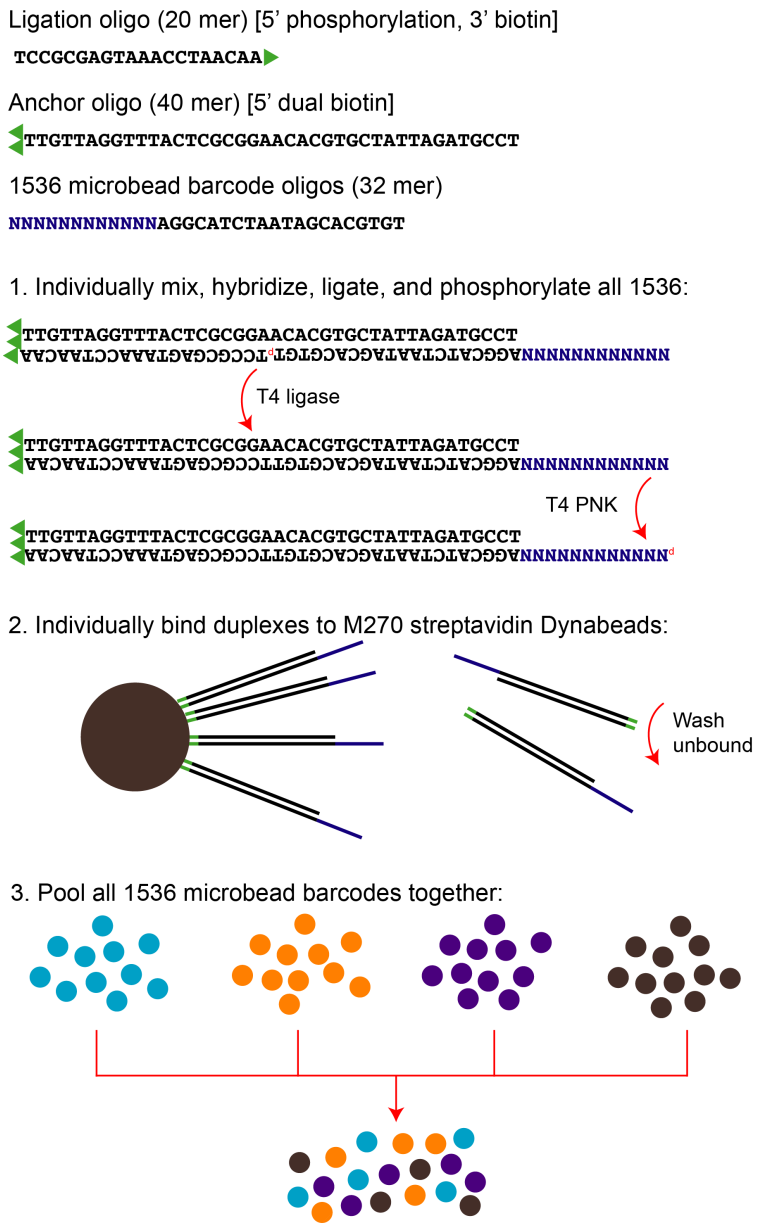

Figure 4.7: **Overview of the 1536-plex barcoded bead generation process.** The 1536-plex barcoded bead generation process is derived from the 384-plex bead generation process originally demonstrated in Plesa et. al2. The process requires 3 oligos: a 20mer ligation oligo with 5' phosphorylation and 3' biotinylation, a 40mer anchor oligo with 5' dual biotinylation, and 1536 32mer microbead barcoded oligos. Each microbead barcoded oligo is individually hybridized to the anchor and ligation oligos in 4 384-well plates, forming three-oligo complexes with 12nt 5' overhangs containing the designed 12mer microbead barcode sequences. T4 ligase then seals the nick between the ligation and microbead barcoded oligo, and T4 PNK phosphorylates the 12nt 5'microbead barcode overhang. All duplexes are then individually bound to M270 Streptavidin Dynabeads, washed, and pooled to form a single 1536-plex barcoded bead pool.

Table 4.1: **The oligos required for the bead barcoding process.** All oligos were ordered from Integrated DNA Technologies.

Oligo Name	Sequence	Modifications	Amount
Ligation oligo	TCCGCGAGTAAACCTAACAA	3' biotin 5' phosphorylation	96.0 nmol
Anchor oligo	TTGTTAGGTTTACTCGCGGAA-CACGTGCTATTAGATGCCT	5' dual biotin	96.0 nmol
384X barcoded oligos	12-mer microbead barcode reverse complement + AGGCATCTAATAGCACGTGT		0.1 nmol each

Table 4.2: **The oligos required for ePCA and single-primer suppression PCR.** The suppression primer aligns to the proximal 20bp of the ITR overhang. All oligos were ordered from Integrated DNA Technologies.

Name	Sequence
AsmF_40bpITR	TAAGCGCCCTTCTAATACCCAGGTCTGGCCCTATATACGAATCGGGGATGGTAACTAACG
AsmR_40bpITR	TAAGCGCCCTTCTAATACCCAGGTCTGGCCCTATATACGAATAGCTGATTGTCCGTTGGT
Suppression primer	AGGTCTGGCCCTATATACGA

Table 4.3: **The primers required to amplify libraries from the OLS pool.** All oligos were ordered from Integrated DNA Technologies.

Library	Codon	AmpF Name	AmpF Sequence	AmpR Name	AmpR Sequence
Control	1	skpp15-9-F	CGATCGTGCCCACCT	skpp15-9-R	GTGCGGGCTCCAACT
Control	2	skpp15-13-F	GGGTTCGAGCGGGAG	skpp15-13-R	GTGCGGGCTCCAACT
Overlap	1	skpp15-23-F	AGCTGCTACACCGCC	skpp15-23-R	GCGCGATGGTCACAG
Overlap	2	skpp15-26-F	GCGGCACCACAAACT	skpp15-26-R	CGTGGCCTCTGTCCT
deltaG	1	skpp15-30-F	TCCACCGTCGGCAAG	skpp15-30-R	GGCCGCACCCAGTAG
deltaG	2	skpp15-33-F	AAGTGCCCTTCCCGT	skpp15-33-R	GAGTCCGCGCAAGAG
BsmAI	1	skpp15-40-F	AGGCGGTCGAGAGTG	skpp15-40-R	CCGTCCTCCACCCAG
BsmAI	2	skpp15-46-F	CCGCATGCAGTCCCT	skpp15-46-R	CGACTCTTGCGCCCT
BsrDI	1	skpp15-49-F	GGCCCAGCGAAGATG	skpp15-49-R	GATCAGCACCGCGAC
BsrDI	2	skpp15-51-F	GGCGCGCTCTAACAC	skpp15-51-R	CTCCCTCTCGCAGCA
1536plex	1	skpp15-56-F	AACGCCCAGCCTGTC	skpp15-56-R	CCGCGTTGCTGAGTG
1536plex	2	skpp15-59-F	AGGCACGCTCAACCT	skpp15-59-R	CCTAGGTCGCACGCA

References

[1] R. T. Hietpas, J. D. Jensen, and D. N. A. Bolon, "Experimental illumination of a fitness landscape," *Proceedings of the National Academy of Sciences*, vol. 108, no. 19, pp. 7896–7901, 2011.

[2] J. B. Kinney, A. Murugan, C. G. Callan, and E. C. Cox, "Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence," *Proceedings of the National Academy of Sciences*, vol. 107, no. 20, pp. 9158–9163, 2010.

[3] E. Sharon, Y. Kalma, A. Sharp, T. Raveh-Sadka, M. Levo, D. Zeevi, L. Keren, Z. Yakhini, A. Weinberger, and E. Segal, "Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters," *Nature Biotechnology*, vol. 30, p. 521, may 2012.

[4] L. M. Starita, D. L. Young, M. Islam, J. O. Kitzman, J. Gullingsrud, R. J. Hause, D. M. Fowler, J. D. Parvin, J. Shendure, and S. Fields, "Massively parallel functional analysis of brca1 ring domain variants," *Genetics*, vol. 200, no. 2, pp. 413–422, 2015.

[5] K. M. Doolan and D. W. Colby, "Conformation-dependent epitopes recognized by prion protein antibodies probed using mutational scanning and deep sequencing," *Journal of Molecular Biology*, vol. 427, no. 2, pp. 328 – 340, 2015.

[6] R. P. Patwardhan, C. Lee, O. Litvin, D. L. Young, D. Pe'er, and J. Shendure, "High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis," *Nature Biotechnology*, vol. 27, pp. 1173–1175, dec 2009.

[7] M. Gasperini, L. Starita, and J. Shendure, "The power of multiplexed functional analysis of genetic variants," *Nature Protocols*, vol. 11, pp. 1782–1787, oct 2016.

[8] C. D. Arnold, D. Gerlach, C. Stelzer, Ł. M. Boryń, M. Rath, and A. Stark, "Genome-wide quantitative enhancer activity maps identified by starr-seq," *Science*, vol. 339, no. 6123, pp. 1074–1077, 2013.

[9] K. S. Sarkisyan, D. A. Bolotin, M. V. Meer, D. R. Usmanova, A. S. Mishin, G. V. Sharonov, D. N. Ivankov, N. G. Bozhanova, M. S. Baranov, O. Soylemez, N. S. Bogatyreva, P. K. Vlasov, E. S. Egorov, M. D. Logacheva, A. S. Kondrashov, D. M. Chudakov, E. V. Putintseva, I. Z. Mamedov, D. S. Tawfik, K. A. Lukyanov, and F. A. Kondrashov, "Local fitness landscape of the green fluorescent protein," *Nature*, vol. 533, pp. 397–401, 2016.

[10] G. J. Rocklin, T. M. Chidyausiku, I. Goreshnik, A. Ford, S. Houliston, A. Lemak, L. Carter, R. Ravichandran, V. K. Mulligan, A. Chevalier, C. H. Arrowsmith, and D. Baker, "Global analysis of protein folding using massively parallel design, synthesis, and testing," *Science*, vol. 357, pp. 168–175, jul 2017.

[11] J. Quan, I. Saaem, N. Tang, S. Ma, N. Negre, H. Gong, K. P. White, and J. Tian, "Parallel on-chip gene synthesis and application to optimization of protein expression," *Nature Biotechnology*, vol. 29, no. 5, pp. 449–452, 2011.

[12] S. Kosuri, N. Eroshenko, E. M. Leproust, M. Super, J. Way, J. B. Li, and G. M. Church, "Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips," *Nature biotechnology*, vol. 28, no. 12, pp. 1295–1299, 2010.

[13] C. Plesa, A. M. Sidore, N. B. Lubock, D. Zhang, and S. Kosuri, "Multiplexed gene synthesis in emulsions for exploring protein functional landscapes," *Science*, vol. 359, pp. 343–347, jan 2018.

[14] J. Smith and P. Modrich, "Removal of polymerase-produced mutant sequences from pcr products," *Proceedings of the National Academy of Sciences*, vol. 94, no. 13, pp. 6847–6850, 1997.

[15] P. A. Carr, J. S. Park, Y.-J. Lee, T. Yu, S. Zhang, and J. M. Jacobson, "Protein-mediated error correction for de novo DNA synthesis.," *Nucleic acids research*, vol. 32, no. 20, p. e162, 2004.

[16] D. A. Shagin, K. A. Lukyanov, L. L. Vagner, and M. V. Matz, "Regulation of average length of complex PCR product.," *Nucleic acids research*, vol. 27, no. 18, p. e23, 1999.

[17] T. H.-C. Hsiau, D. Sukovich, P. Elms, R. N. Prince, T. Stritmatter, P. Ruan, B. Curry, P. Anderson, J. Sampson, and J. C. Anderson, "A method for multiplex gene synthesis employing error correction based on expression," *PLOS ONE*, vol. 10, pp. 1–15, 03 2015.

[18] N. B. Lubock, D. Zhang, A. M. Sidore, G. M. Church, and S. Kosuri, "A systematic comparison of error correction enzymes by next-generation sequencing," *Nucleic acids research*, vol. 45, no. 15, pp. 9206–9217, 2017.

[19] N. Eroshenko, S. Kosuri, A. H. Marblestone, N. Conway, and G. M. Church, "Gene assembly from chip-synthesized oligonucleotides," *Current Protocols in Chemical Biology*, vol. 4, no. 1, pp. 1–17, 2012.

[20] "Activity of restriction enzymes in PCR buffers." `neb.com/tools-and-resources/usage-guidelines/activity-of-restriction-enzymes-in-pcr-buffers`.

[21] J. J. Schwartz, C. Lee, and J. Shendure, "Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules.," *Nature methods*, vol. 9, no. 9, pp. 913–5, 2012.

[22] W. J. Kent, "Blat—the blast-like alignment tool," *Genome Research*, vol. 12, no. 4, pp. 656–664, 2002.

Chapter 5

Conclusion

5.1 Summary of Novel Technology

This dissertation describes an improved method for the multiplexed synthesis of gene libraries. By compartmentalizing and assembling microarray-derived oligos in emulsions, we significantly reduce the cost and effort of gene synthesis. We further show that these gene libraries can be directly inputted into multiplexed functional assays. Broadly, this technology can be used to learn sequence-function relationships from thousands of synthetic long-length DNA sequences.

5.2 Summary of Findings

In Chapter 2, we accurately quantify error rates in model gene assemblies using next-generation sequencing. We use this experimental and computational pipeline to systematically compare the effects of two polymerases and several enzymatic error correction methods on model gene assemblies. This pipeline allows us to generate conclusions about particular error correction enzymes and polymerases used in gene synthesis protocols. For instance, we find that MutS is the preferred error correction enzyme for increasing the number of perfect assemblies, while ErrASE is more effective at decreasing average error rates. Furthermore, we find that KAPA2G Robust has a significantly higher mismatch rate than NEB Q5, a high fidelity polymerase.

In Chapter 3, we introduce DropSynth, a low-cost, multiplexed method for the construction of gene libraries. This method assembles genes by hybridizing microarray-derived oligos to barcoded

beads, compartmentalizing the bound beads in emulsions, and assembling the contributing oligos into full-length genes. We show that this method can assemble gene libraries ranging from 1,000 to 10,000 constructs at a high rate of representation. Furthermore, we demonstrate that these gene libraries can be directly inputted into multiplexed functional assays. In particular, we assemble over one thousand homologs of PPAT, an essential enzyme in E. coli, and perform a pooled complementation assay where we observe the ability of all homologs to complement native E. coli function. In addition, mismatches generated in the gene assembly allow us to explore the local fitness landscape around each homolog, generating a broad mutational scanning data set. Taken as a whole, this chapter reveals the power of multiplexed gene synthesis for the analysis of sequence-function relationships.

In Chapter 4, we optimize and improve DropSynth, resulting in significant enhancements in the fidelity and scalability of gene assemblies. In particular, we build upon knowledge gained in Chapter 2 by incorporating both enzymatic error correction and high fidelity polymerases into our workflow. These incorporations allow us to improve the fidelity of assembled genes several-fold. Furthermore, we scale up our barcoded bead pool, allowing for the one-pot assembly of over 1,500 designed genes. By improving both fidelity and scalability, we show that DropSynth can be used to generate longer and larger DNA libraries. These optimizations enable the design of multiplexed functional assays with fewer restrictions on the size, length, and content of libraries.

5.3 Future Directions

Our initial demonstrations of DropSynth have generated thousands of protein-coding genes ranging in size from 381-669 bp. Though this is useful, the utility of DropSynth can be spread far beyond what we have demonstrated. For instance, though DropSynth has been validated on genes of up to 530bp in length, the median length of bacterial proteins is nearly 900bp [1]. While we have verified gene assemblies of up to 669 bp on an agarose gel, we currently lack the ability to validate assemblies longer than 500bp due to read length limitations of Illumina sequencing. In order to sequence and characterize longer gene libraries, alternative sequencing technologies are necessary. Improvements in long-read sequencing technologies such as PacBio [2] or Oxford Nanopore Technologies [3] would enable error characterization of long-length genes. Unfortunately, the error

rates of both technologies make accurate assembly characterization difficult [4, 5]. Alternatively, strategies such as Intramolecular-ligated Nanopore Consensus Sequencing (INC-Seq) [6] could be used to circularize the assembled genes, amplify using rolling circle amplification, shear, size-select, and sequence using an Oxford Nanopore MinION sequencer. This strategy allows for the sequencing of many copies of each variant, thus overcoming high error rates associated with MinION sequencing.

In addition, generating longer genes necessitates the stitching together of more oligos. This results in a corresponding decrease in the fidelity of the resulting assemblies, since the fidelity of genes is strongly affected by the fidelity of the oligos. Alternatively, using longer oligos would allow longer genes to be assembled with higher fidelity. Although we have demonstrated assemblies using 230nt oligos from Agilent, 300mer microarrays are currently in development by several companies. Such lengths would allow genes >1 kb to be assembled using only five oligos, which we have already demonstrated is possible.

Up until now, DropSynth has only been used to build protein-coding gene libraries. Though this is useful, a large number of biological questions remain about non-coding regions of the genome, including promoters, splice sites, and regulatory regions [7, 8]. With our demonstrated improvements in assembly fidelity and scalability, we can now build large libraries of long >500bp non-coding regions of the genome. The multiplexed construction of these regions accompanied with a massively parallel reporter assay [9] would result in the large-scale analysis of thousands of long regulatory regions.

Though we have optimized DropSynth extensively, there are still several avenues for improvement. In particular, it would be of considerable use to identify sequence-specific motifs in the oligo design that contribute to assembly failure. These motifs could then be selected against in future designs, resulting in significantly improved library representation. In addition, a large-scale method for isolating individual genes, such as dial-out PCR [10] would be useful for researchers or commercial partners who want to isolate and purify individual genes.

DropSynth is a living protocol, and future updates and improvements will be compiled on https://www.dropsynth.org. In addition, the website contains all necessary protocols and software to replicate the existing technique. As the protocol evolves, we believe future improvements will dramatically accelerate our ability to interrogate sequence-function relationships.

References

[1] L. Brocchieri and S. Karlin, "Protein length in eukaryotic and prokaryotic proteomes," *Nucleic Acids Res.*, vol. 33, pp. 3390–3400, June 2005.

[2] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. deWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, "Real-time dna sequencing from single polymerase molecules," *Science*, vol. 323, no. 5910, pp. 133–138, 2009.

[3] S. Howorka, S. Cheley, and H. Bayley, "Sequence-specific detection of individual DNA strands using engineered nanopores," *Nat. Biotechnol.*, vol. 19, pp. 636–639, July 2001.

[4] M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu, "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers," *BMC Genomics*, vol. 13, no. 1, p. 341, 2012.

[5] A. S. Mikheyev and M. M. Y. Tin, "A first look at the oxford nanopore minion sequencer," *Molecular Ecology Resources*, vol. 14, no. 6, pp. 1097–1102, 2014.

[6] C. Li, K. R. Chng, E. J. H. Boey, A. H. Q. Ng, A. Wilm, and N. Nagarajan, "INC-Seq: accurate single molecule reads using nanopore sequencing," *Gigascience*, vol. 5, p. 34, Aug. 2016.

[7] M. Gasperini, L. Starita, and J. Shendure, "The power of multiplexed functional analysis of genetic variants," *Nat. Protoc.*, vol. 11, pp. 1782–1787, Oct. 2016.

[8] J. Weile and F. P. Roth, "Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas," *Human Genetics*, vol. 137, pp. 665–678, Sep 2018.

[9] F. Inoue and N. Ahituv, "Decoding enhancers using massively parallel reporter assays," *Genomics*, vol. 106, pp. 159–164, Sept. 2015.

[10] J. J. Schwartz, C. Lee, and J. Shendure, "Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules," *Nat. Methods*, vol. 9, pp. 913–915, Sept. 2012.